Time for a Change: Exploring New Avenues for Email Collection Preservation

Smithsonian
Libraries and Archives

Lynda Schmitz Fuhrig

Digital Archivist

Email Archiving Symposium
June 15, 2023

Through the Years

Elm: IMAP folder hurtta@posti.fmi.fi:RoskaPosti - Shell - Konsole
Session Edit View Settings Help
10 USERS
GROUP PACK
Lotus
CC:MAIL
E-Mail and scheduling for Windows

```
            IMAP folder is 'hurtta@posti.fmi.fi:RoskaPosti' with 789 messages
                                [ELM 2.4ME+ PL116 (25)]

  OM 1   May 19 Travis@menws12.com (70)   less expensive & No pres~cription F E E S
  O! 2   May 19 rncuka@yahoo.com   (?)    URGENT CONTACT                           yvmk
  OM 3   May 19 Thacker            (71)   Karaila What meds do u need?
  OM 4   May 19 Bernadine F. Ryan  (63)   Buy Vicodin at Unbelievable Prices
  OM 5   May 19 Erik Weiss         (55)   Buy Vicodin Pain Pills Now US Doctors US Phar
  OM 6   May 19 Lakeisha Sawyer    (63)   unb_eatable software dea-ls
  O! 7   May 19 Earnestine Seymour (?)    Top Quality
  OM 8   May 19 Angelique Moon     (85)   Re:Real Estate Millions evil
  OM 9   May 19 Gino Dill          (83)   Re:Notice of Default heighten
  OM 10  May 19 Rae Kinney         (53)   wormy
  OM 11  May 19 Aron Mccauley      (54)   melee
  OM 12  May 18 Jenna Hart         (62)   stoneware Searching for CHE@P high-quality M!
  OM 13  May 19 Inez T. Turner     (60)   Free Vicodin Prescriptions.
  O! 14  May 19 Bjarne AnderBlanke (?)    Re: Re: Your winning ebay auction
  OM 15  May 19 Mohamed Black      (46)   limite.d time prices
  OM 16  May 18 Comer, Velma       (64)   XOANAX, VALIU.M , Cia+lis, VIAG2RA and many m
  O! 17  May 18 Ulysses Travis     (?)    Major Gains-A Story Stock Steal? J k  RND_LC_


           !=pipe, !=shell, ?=help, <n>=set current to n, /=search pattern
       a)lias, C)opy, c)hange folder, d)elete, e)dit, f)orward, g)roup reply, m)ail,
        n)ext, o)ptions, p)rint, q)uit, r)eply, s)ave, t)ag, u)ndelete, or e(x)it

Command: □
```

Lotus Notes

Groupware and E-mail for the Net

Novell.

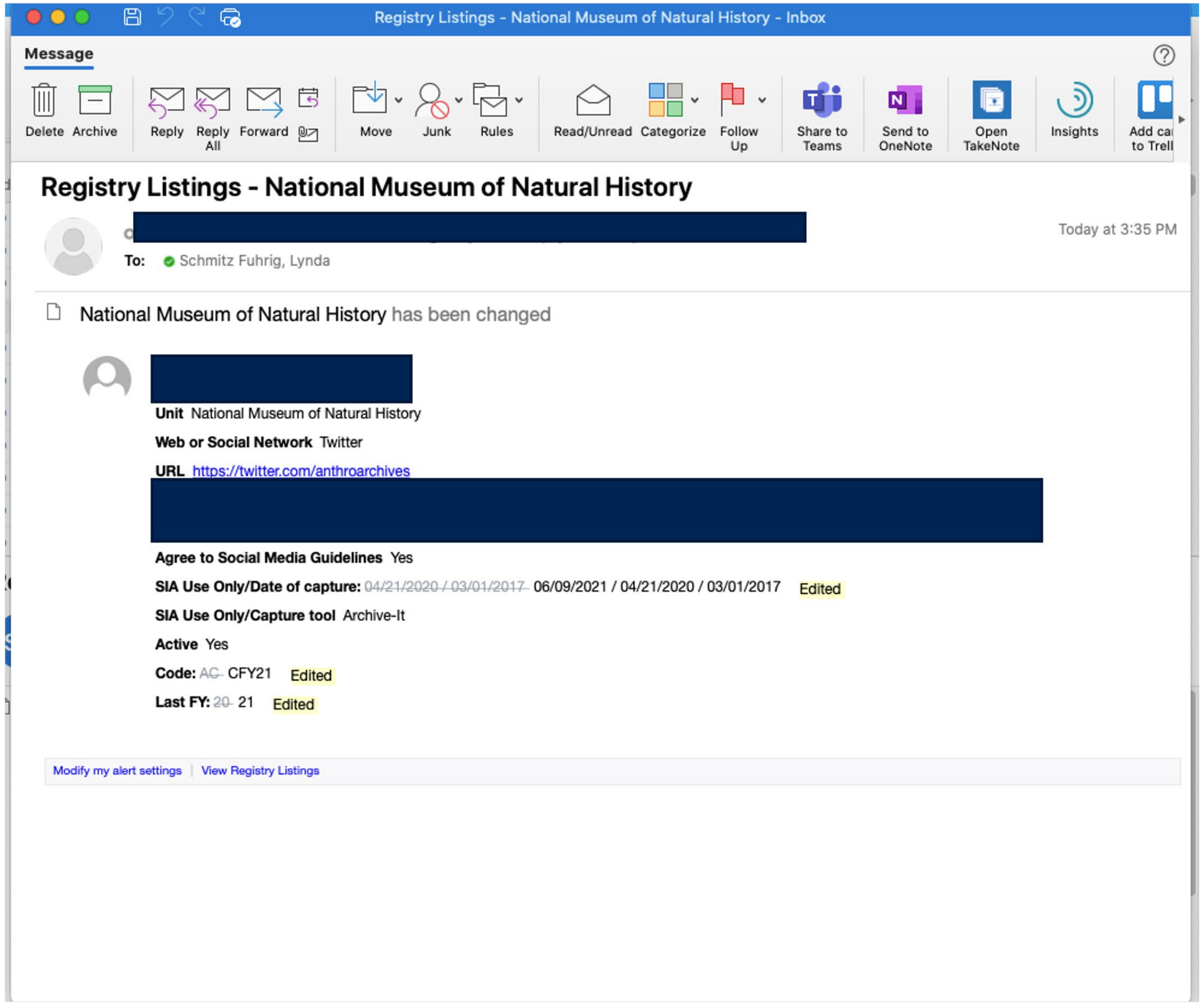
GroupWise™

Outlook® ™

Smithsonian
Libraries and Archives

Accessioned accounts

- Role at the Smithsonian or specific project – Secretary, undersecretaries, directors, curators, etc.
- When the account holder leaves the Smithsonian, but can be taken in sooner
- No weeding done by SIA except Junk Email or Deleted Items folders in some cases
- 15-year-access restriction. PII permanently restricted



Registry Listings - National Museum of Natural History

[redacted] Today at 3:35 PM

To: Schmitz Fuhrig, Lynda

National Museum of Natural History has been changed

[redacted]

Unit National Museum of Natural History

Web or Social Network Twitter

URL https://twitter.com/anthroarchives

[redacted]

Agree to Social Media Guidelines Yes

SIA Use Only/Date of capture: ~~04/21/2020 / 03/01/2017~~ 06/09/2021 / 04/21/2020 / 03/01/2017 Edited

SIA Use Only/Capture tool Archive-It

Active Yes

Code: ~~AC~~ CFY21 Edited

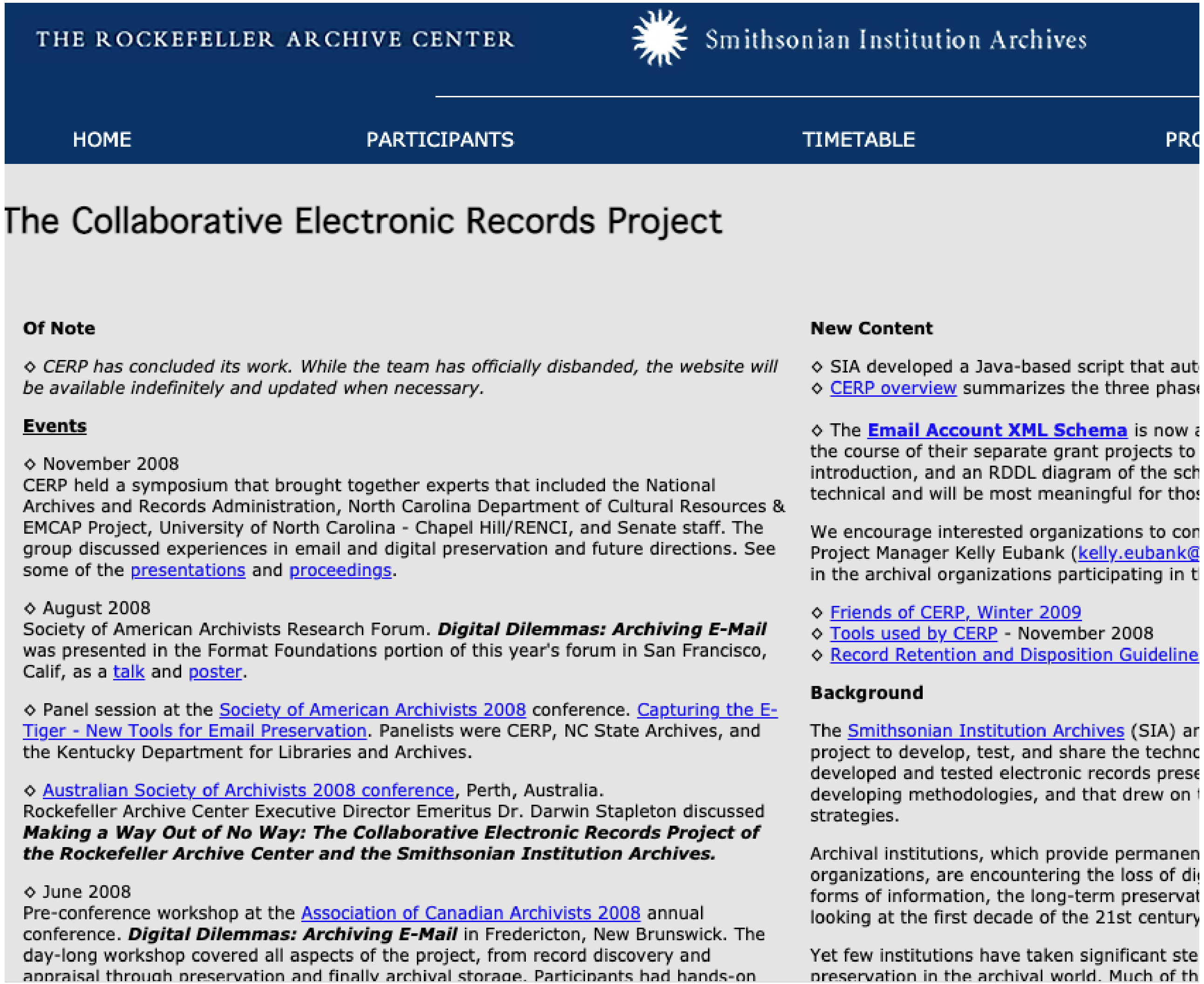
Last FY: ~~20~~ 21 Edited

Modify my alert settings | View Registry Listings

Smithsonian
Libraries and Archives

CERP in 2005

- Smithsonian Institution Archives and Rockefeller Archives Center pilot
- Three Smithsonian offices
- Practices with both email and digital records
- SI migrating from GroupWise to Outlook
- Tested various tools and worked with another project at the State Archives of North Carolina, KY, and PA - EMCAP


THE ROCKEFELLER ARCHIVE CENTER

Smithsonian Institution Archives

The Collaborative Electronic Records Project

Of Note

◇ *CERP has concluded its work. While the team has officially disbanded, the website will be available indefinitely and updated when necessary.*

Events

◇ November 2008
CERP held a symposium that brought together experts that included the National Archives and Records Administration, North Carolina Department of Cultural Resources & EMCAP Project, University of North Carolina - Chapel Hill/RENCI, and Senate staff. The group discussed experiences in email and digital preservation and future directions. See some of the presentations and proceedings.

◇ August 2008
Society of American Archivists Research Forum. ***Digital Dilemmas: Archiving E-Mail*** was presented in the Format Foundations portion of this year's forum in San Francisco, Calif, as a talk and poster.

◇ Panel session at the Society of American Archivists 2008 conference. Capturing the E-Tiger - New Tools for Email Preservation. Panelists were CERP, NC State Archives, and the Kentucky Department for Libraries and Archives.

◇ Australian Society of Archivists 2008 conference, Perth, Australia.
Rockefeller Archive Center Executive Director Emeritus Dr. Darwin Stapleton discussed ***Making a Way Out of No Way: The Collaborative Electronic Records Project of the Rockefeller Archive Center and the Smithsonian Institution Archives.***

◇ June 2008
Pre-conference workshop at the Association of Canadian Archivists 2008 annual conference. ***Digital Dilemmas: Archiving E-Mail*** in Fredericton, New Brunswick. The day-long workshop covered all aspects of the project, from record discovery and appraisal through preservation and finally archival storage. Participants had hands-on

New Content

◇ SIA developed a Java-based script that aut
◇ CERP overview summarizes the three phas

◇ The **Email Account XML Schema** is now a
the course of their separate grant projects to
introduction, and an RDDL diagram of the sch
technical and will be most meaningful for thos

We encourage interested organizations to con
Project Manager Kelly Eubank (kelly.eubank@
in the archival organizations participating in t

◇ Friends of CERP, Winter 2009
◇ Tools used by CERP - November 2008
◇ Record Retention and Disposition Guideline

Background

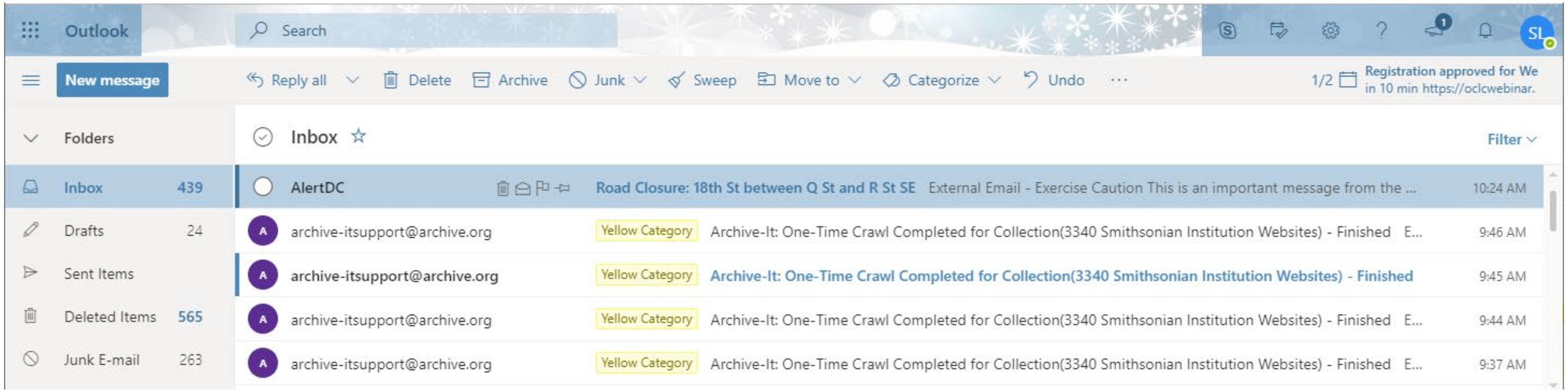
The Smithsonian Institution Archives (SIA) ar
project to develop, test, and share the techno
developed and tested electronic records prese
developing methodologies, and that drew on t
strategies.

Archival institutions, which provide permanen
organizations, are encountering the loss of di
forms of information, the long-term preservat
looking at the first decade of the 21st century

Yet few institutions have taken significant ste
preservation in the archival world. Much of th

Smithsonian
Libraries and Archives

Our Approach



Registration approved for We in 10 min https://oclcwebinar.

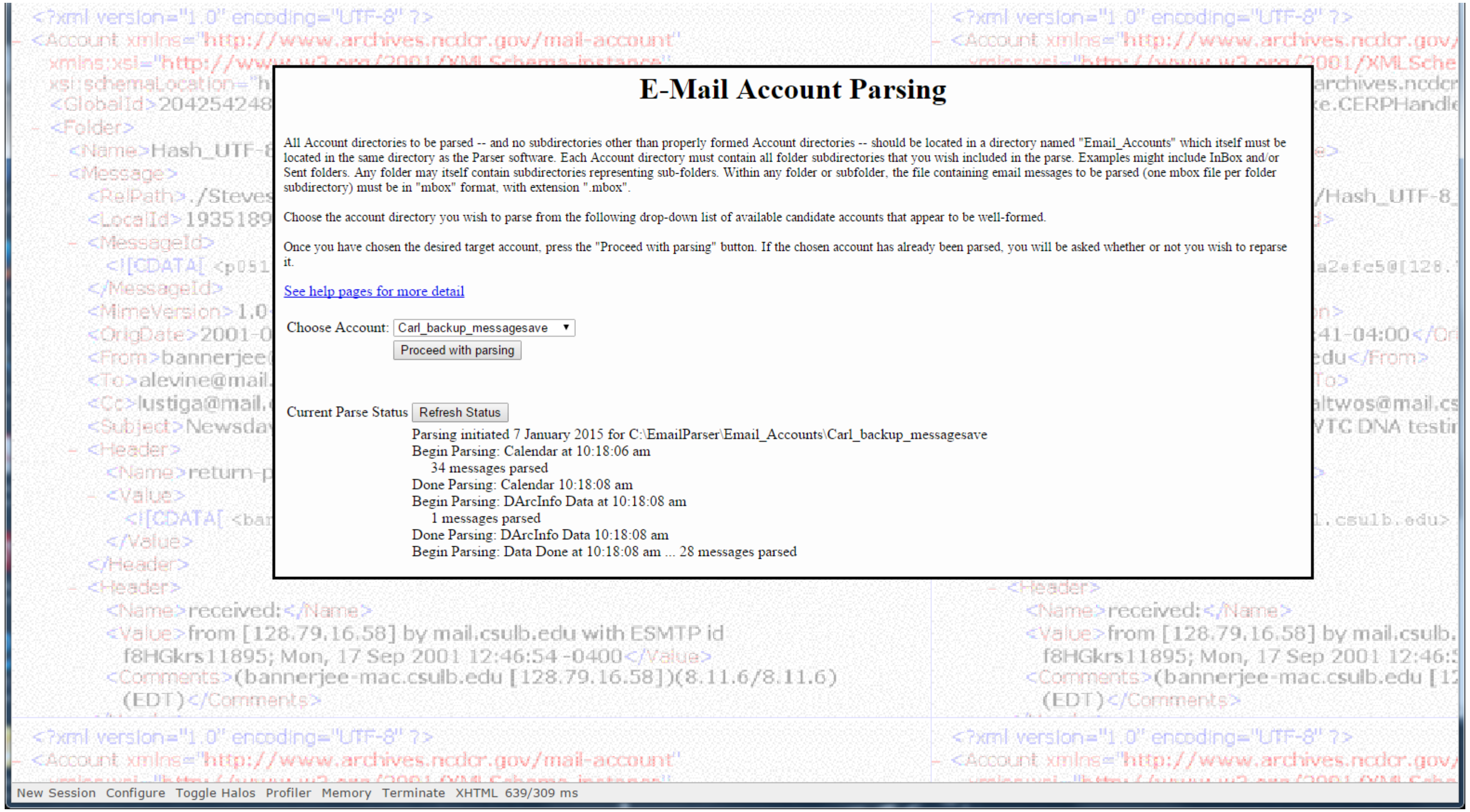
AlertDC		Road Closure: 18th St between Q St and R St SE External Email - Exercise Caution This is an important message from the ...	10:24 AM
archive-itsupport@archive.org	Yellow Category	Archive-It: One-Time Crawl Completed for Collection(3340 Smithsonian Institution Websites) - Finished E...	9:46 AM
archive-itsupport@archive.org	Yellow Category	Archive-It: One-Time Crawl Completed for Collection(3340 Smithsonian Institution Websites) - Finished	9:45 AM
archive-itsupport@archive.org	Yellow Category	Archive-It: One-Time Crawl Completed for Collection(3340 Smithsonian Institution Websites) - Finished E...	9:44 AM
archive-itsupport@archive.org	Yellow Category	Archive-It: One-Time Crawl Completed for Collection(3340 Smithsonian Institution Websites) - Finished E...	9:37 AM

- Open formats
- Anyone could adopt/adapt
- Accounts, not individual emails
  - Easier manageability at this level

Smithsonian
Libraries and Archives

CERP Parser

Powered by Squeak



E-Mail Account Parsing

All Account directories to be parsed -- and no subdirectories other than properly formed Account directories -- should be located in a directory named "Email_Accounts" which itself must be located in the same directory as the Parser software. Each Account directory must contain all folder subdirectories that you wish included in the parse. Examples might include InBox and/or Sent folders. Any folder may itself contain subdirectories representing sub-folders. Within any folder or subfolder, the file containing email messages to be parsed (one mbox file per folder subdirectory) must be in "mbox" format, with extension ".mbox".

Choose the account directory you wish to parse from the following drop-down list of available candidate accounts that appear to be well-formed.

Once you have chosen the desired target account, press the "Proceed with parsing" button. If the chosen account has already been parsed, you will be asked whether or not you wish to reparse it.

See help pages for more detail

Choose Account: Carl_backup_messagesave

Proceed with parsing
Refresh Status
New Session Configure Toggle Halos Profiler Memory Terminate XHTML 639/309 ms

Current Parse Status

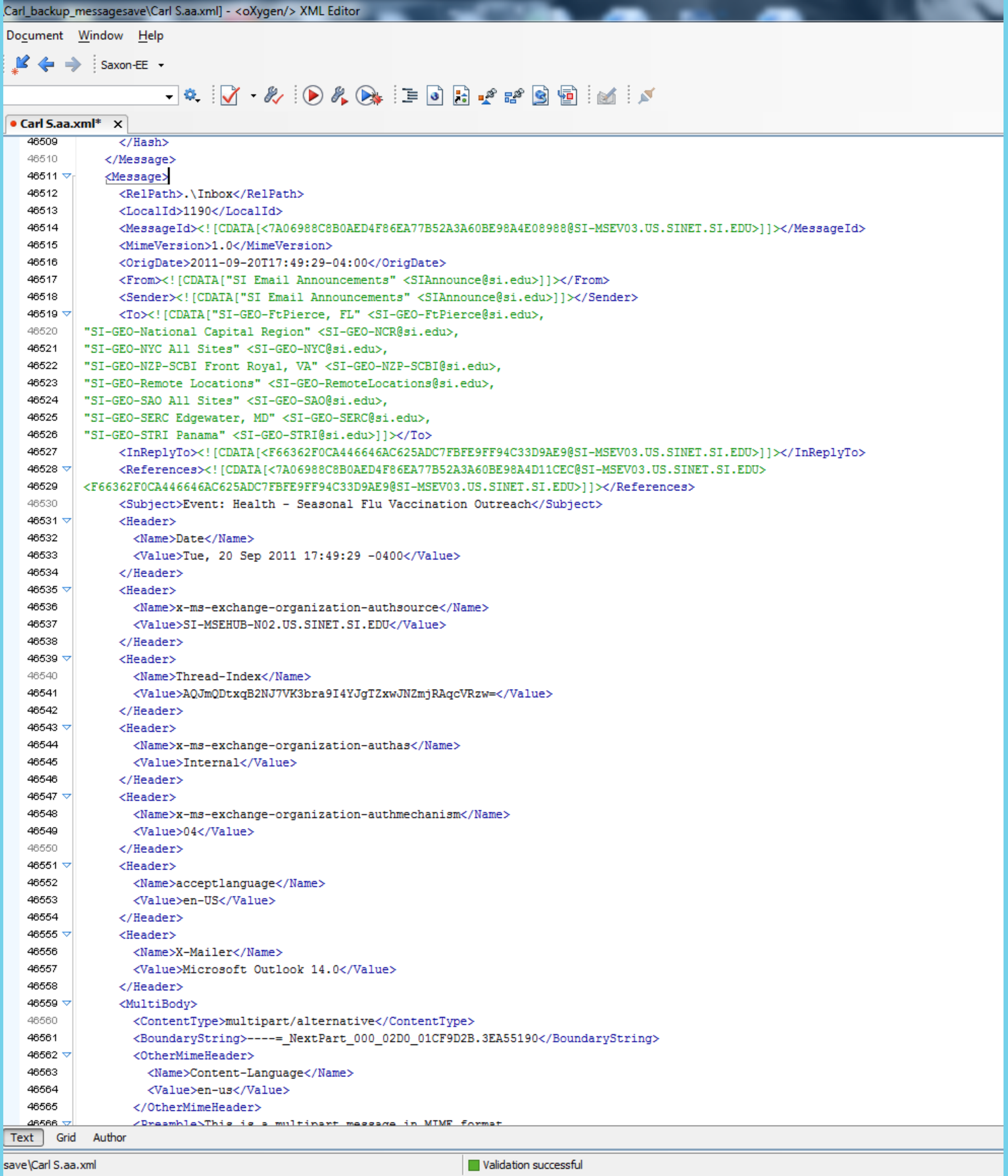
```
Parsing initiated 7 January 2015 for C:\EmailParser\Email_Accounts\Carl_backup_messagesave
Begin Parsing: Calendar at 10:18:06 am
    34 messages parsed
Done Parsing: Calendar 10:18:08 am
Begin Parsing: DArcInfo Data at 10:18:08 am
    1 messages parsed
Done Parsing: DArcInfo Data 10:18:08 am
Begin Parsing: Data Done at 10:18:08 am ... 28 messages parsed
```

Smithsonian
Libraries and Archives

Preservation

- Preservation output through Email Account XML Schema (EAXS) using the parser

- Human and machine readable but not “pretty”



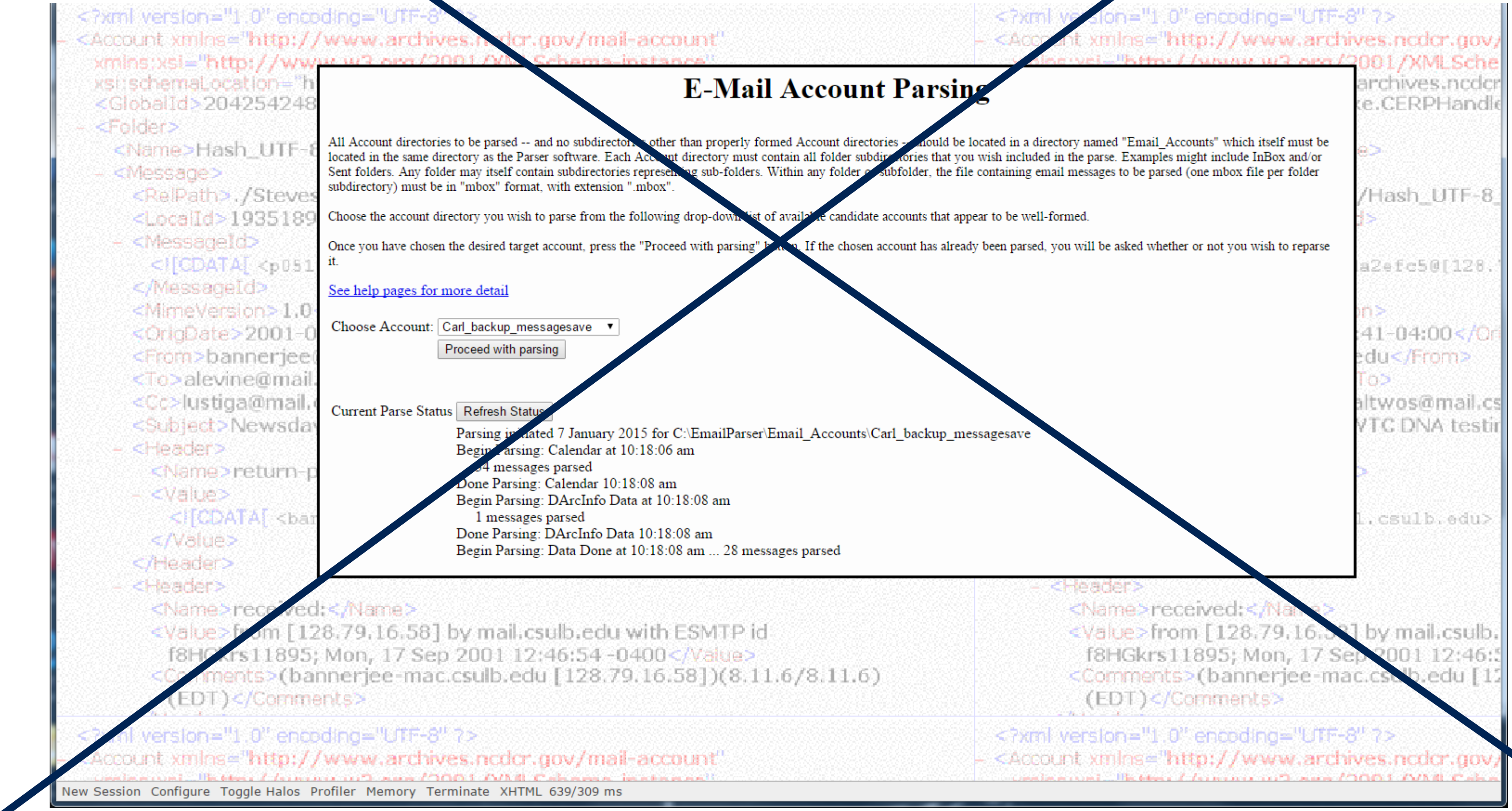
```
    </Hash>
  </Message>
  <Message>
    <RelPath>.\Inbox</RelPath>
    <LocalId>1190</LocalId>
    <MessageId><![CDATA[<7A06988C8B0AED4F86EA77B52A3A60BE98A4E08988@SI-MSEV03.US.SINET.SI.EDU>]]></MessageId>
    <MimeVersion>1.0</MimeVersion>
    <OrigDate>2011-09-20T17:49:29-04:00</OrigDate>
    <From><![CDATA["SI Email Announcements" <SIAnnounce@si.edu>]]></From>
    <Sender><![CDATA["SI Email Announcements" <SIAnnounce@si.edu>]]></Sender>
    <To><![CDATA["SI-GEO-FtPierce, FL" <SI-GEO-FtPierce@si.edu>,
"SI-GEO-National Capital Region" <SI-GEO-NCR@si.edu>,
"SI-GEO-NYC All Sites" <SI-GEO-NYC@si.edu>,
"SI-GEO-NZP-SCBI Front Royal, VA" <SI-GEO-NZP-SCBI@si.edu>,
"SI-GEO-Remote Locations" <SI-GEO-RemoteLocations@si.edu>,
"SI-GEO-SAO All Sites" <SI-GEO-SAO@si.edu>,
"SI-GEO-SERC Edgewater, MD" <SI-GEO-SERC@si.edu>,
"SI-GEO-STRI Panama" <SI-GEO-STRI@si.edu>]]></To>
    <InReplyTo><![CDATA[<F66362F0CA446646AC625ADC7FBFE9FF94C33D9AE9@SI-MSEV03.US.SINET.SI.EDU>]]></InReplyTo>
    <References><![CDATA[<7A06988C8B0AED4F86EA77B52A3A60BE98A4D11CEC@SI-MSEV03.US.SINET.SI.EDU>
<F66362F0CA446646AC625ADC7FBFE9FF94C33D9AE9@SI-MSEV03.US.SINET.SI.EDU>]]></References>
    <Subject>Event: Health - Seasonal Flu Vaccination Outreach</Subject>
    <Header>
      <Name>Date</Name>
      <Value>Tue, 20 Sep 2011 17:49:29 -0400</Value>
    </Header>
    <Header>
      <Name>x-ms-exchange-organization-authsource</Name>
      <Value>SI-MSEHUB-N02.US.SINET.SI.EDU</Value>
    </Header>
    <Header>
      <Name>Thread-Index</Name>
      <Value>AQJmQDtxqB2NJ7VK3bra9I4YJgTZxwJNZmjRAqcVRzw=</Value>
    </Header>
    <Header>
      <Name>x-ms-exchange-organization-authas</Name>
      <Value>Internal</Value>
    </Header>
    <Header>
      <Name>x-ms-exchange-organization-authmechanism</Name>
      <Value>04</Value>
    </Header>
    <Header>
      <Name>acceptlanguage</Name>
      <Value>en-US</Value>
    </Header>
    <Header>
      <Name>X-Mailer</Name>
      <Value>Microsoft Outlook 14.0</Value>
    </Header>
    <MultiBody>
      <ContentType>multipart/alternative</ContentType>
      <BoundaryString>----=_NextPart_000_02D0_01CF9D2B.3EA55190</BoundaryString>
      <OtherMimeHeader>
        <Name>Content-Language</Name>
        <Value>en-us</Value>
      </OtherMimeHeader>
      <Preamble>This is a multipart message in MIME format.
```

Validation successful

Smithsonian
Libraries and Archives

CERP Parser

Powered by Squeak



E-Mail Account Parsing

All Account directories to be parsed -- and no subdirectories other than properly formed Account directories -- should be located in a directory named "Email_Accounts" which itself must be located in the same directory as the Parser software. Each Account directory must contain all folder subdirectories that you wish included in the parse. Examples might include InBox and/or Sent folders. Any folder may itself contain subdirectories representing sub-folders. Within any folder or subfolder, the file containing email messages to be parsed (one mbox file per folder subdirectory) must be in "mbox" format, with extension ".mbox".

Choose the account directory you wish to parse from the following drop-down list of available candidate accounts that appear to be well-formed.

Once you have chosen the desired target account, press the "Proceed with parsing" button. If the chosen account has already been parsed, you will be asked whether or not you wish to reparse it.

See help pages for more detail

Choose Account: Carl_backup_messagesave ▾

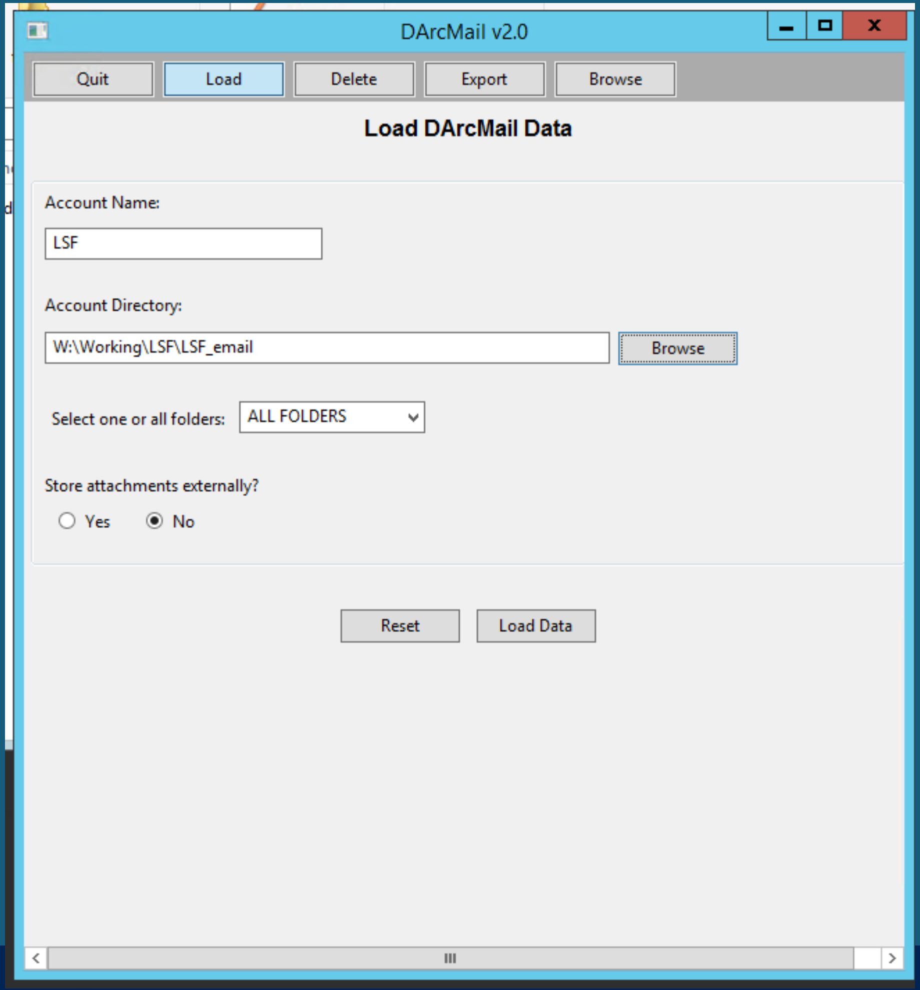
Current Parse Status

```
Parsing initiated 7 January 2015 for C:\EmailParser\Email_Accounts\Carl_backup_messagesave
Begin Parsing: Calendar at 10:18:06 am
   34 messages parsed
Done Parsing: Calendar 10:18:08 am
Begin Parsing: DArcInfo Data at 10:18:08 am
   1 messages parsed
Done Parsing: DArcInfo Data 10:18:08 am
Begin Parsing: Data Done at 10:18:08 am ... 28 messages parsed
```

Proceed with parsing
Refresh Status
New Session Configure Toggle Halos Profiler Memory Terminate XHTML 639/309 ms

Smithsonian
Libraries and Archives

DArcMail (Digital Archive for eMail) Suite



Load DArcMail Data

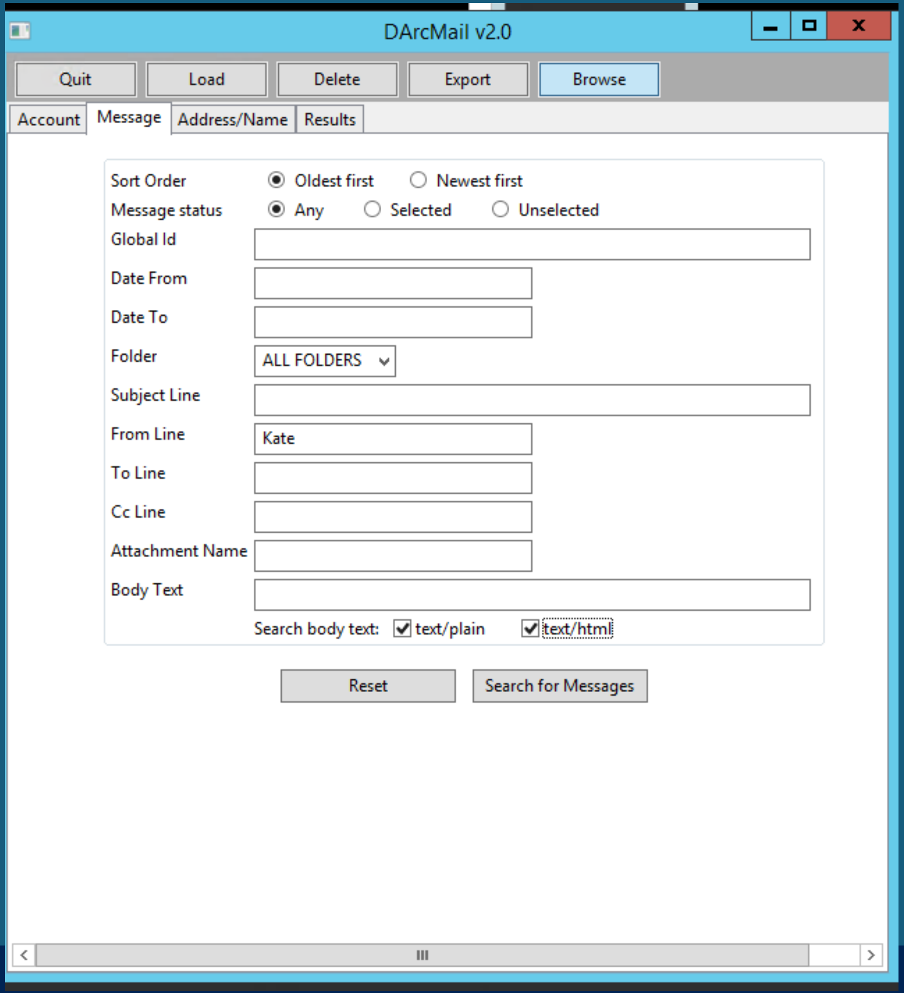
Account Name: LSF

Account Directory: W:\Working\LSF\LSF_email

Select one or all folders: ALL FOLDERS

Store attachments externally?

○ Yes ● No



Sort Order ● Oldest first ○ Newest first

Message status ● Any ○ Selected ○ Unselected

Global Id

Date From

Date To

Folder ALL FOLDERS

Subject Line

From Line Kate

To Line

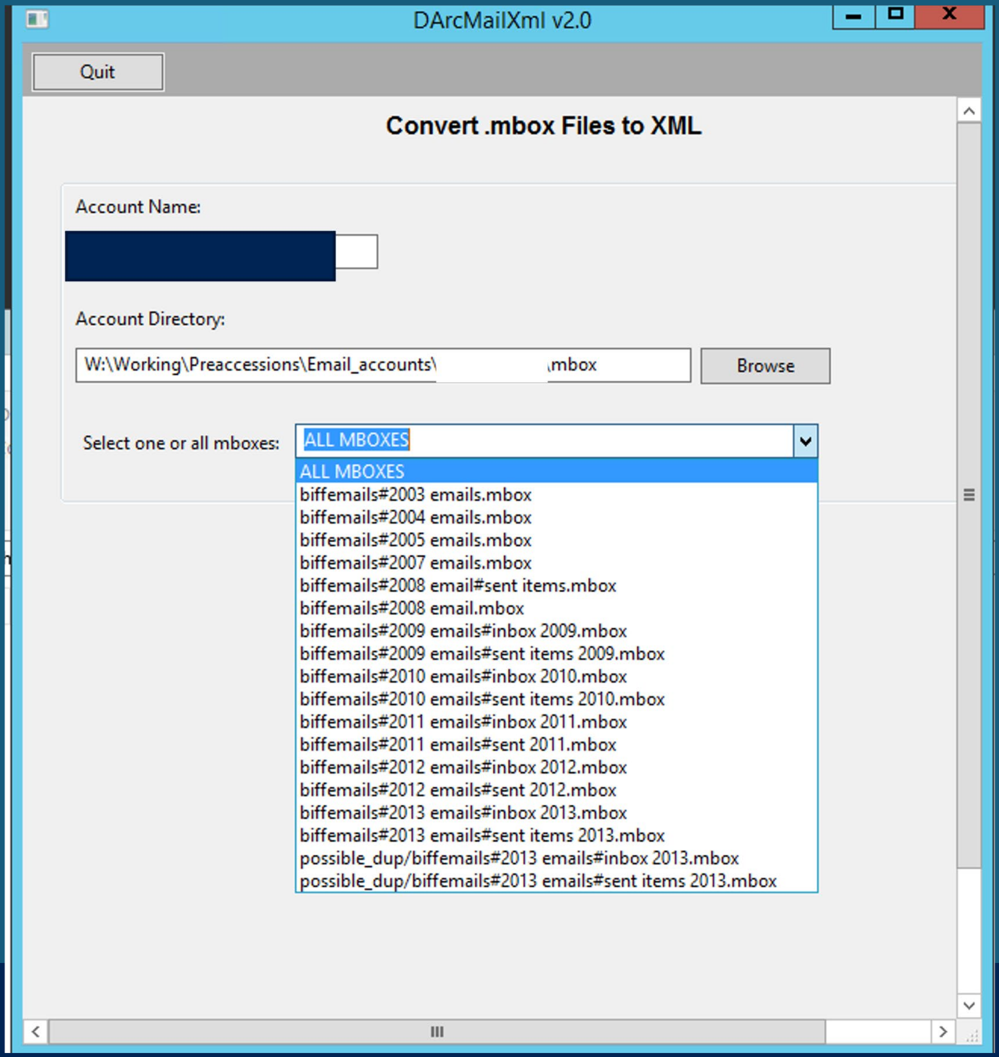
Cc Line

Attachment Name

Body Text

Search body text: ☑ text/plain ☑ text/html

DarcMailXml conversion from MBOX to XML



Convert .mbox Files to XML

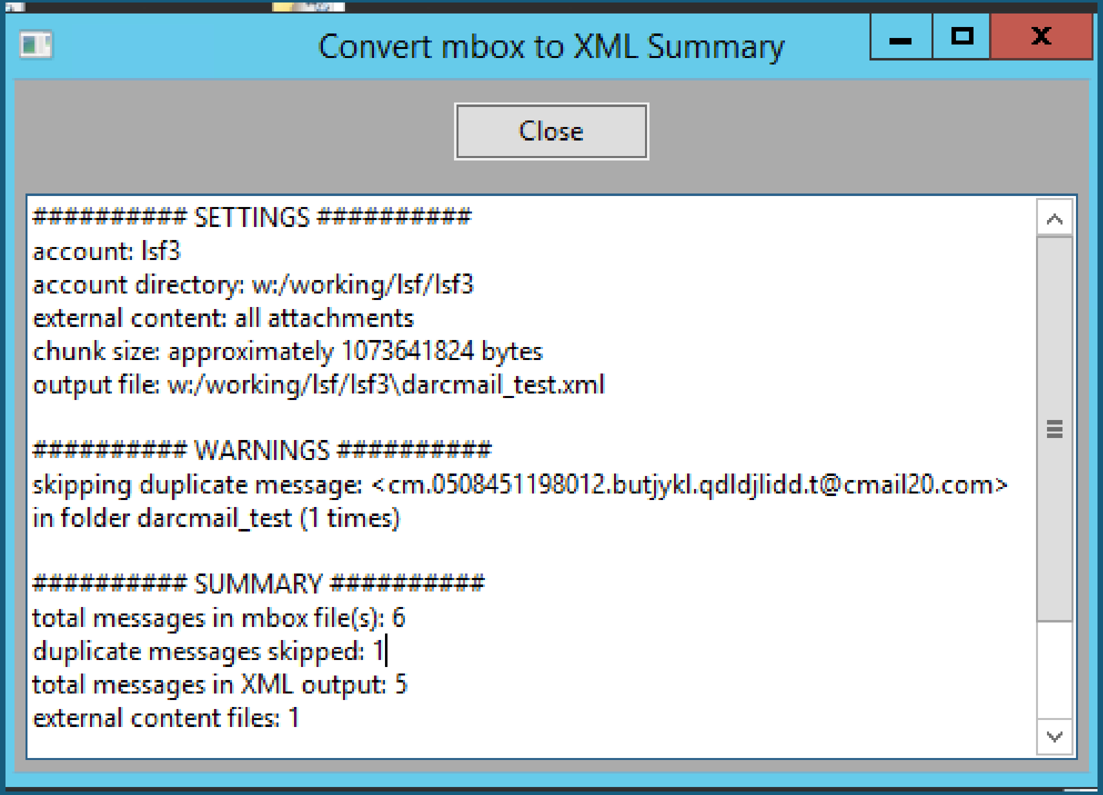
Account Name: [redacted]

Account Directory: W:\Working\Preaccessions\Email_accounts\[redacted].mbox

Select one or all mboxes: ALL MBOXES

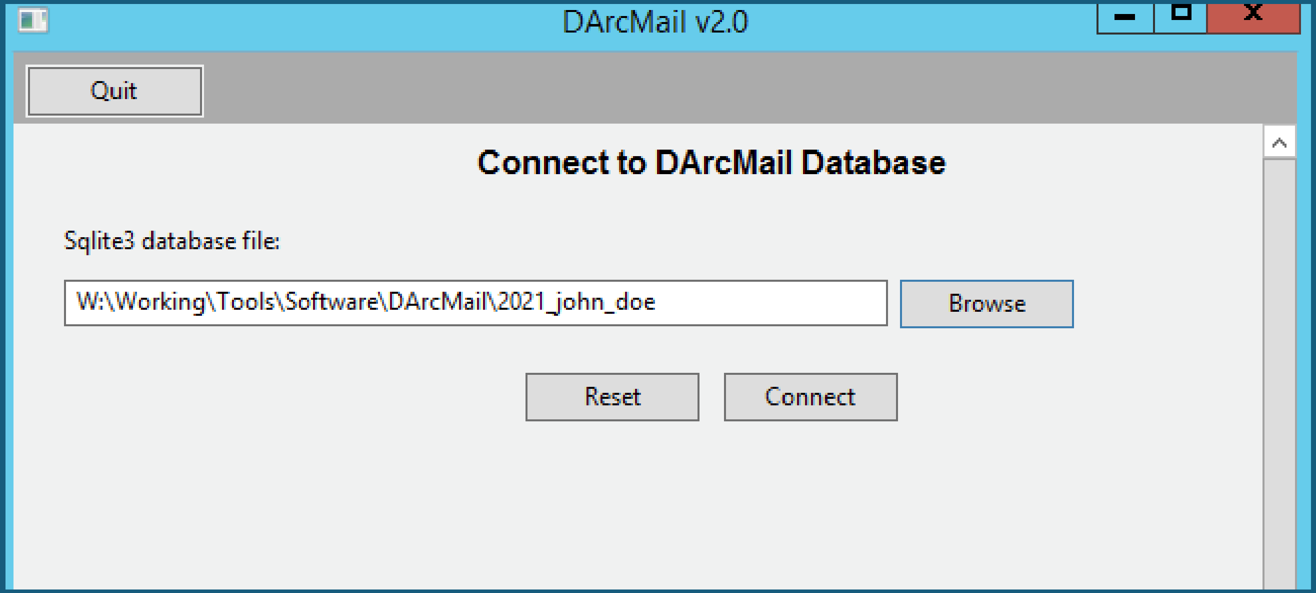
ALL MBOXES
biffemails#2003 emails.mbox
biffemails#2004 emails.mbox
biffemails#2005 emails.mbox
biffemails#2007 emails.mbox
biffemails#2008 email#sent items.mbox
biffemails#2008 email.mbox
biffemails#2009 emails#inbox 2009.mbox
biffemails#2009 emails#sent items 2009.mbox
biffemails#2010 emails#inbox 2010.mbox
biffemails#2010 emails#sent items 2010.mbox
biffemails#2011 emails#inbox 2011.mbox
biffemails#2011 emails#sent 2011.mbox
biffemails#2012 emails#inbox 2012.mbox
biffemails#2012 emails#sent 2012.mbox
biffemails#2013 emails#inbox 2013.mbox
biffemails#2013 emails#sent items 2013.mbox
possible_dup/biffemails#2013 emails#inbox 2013.mbox
possible_dup/biffemails#2013 emails#sent items 2013.mbox

Log for conversion to preservation XML



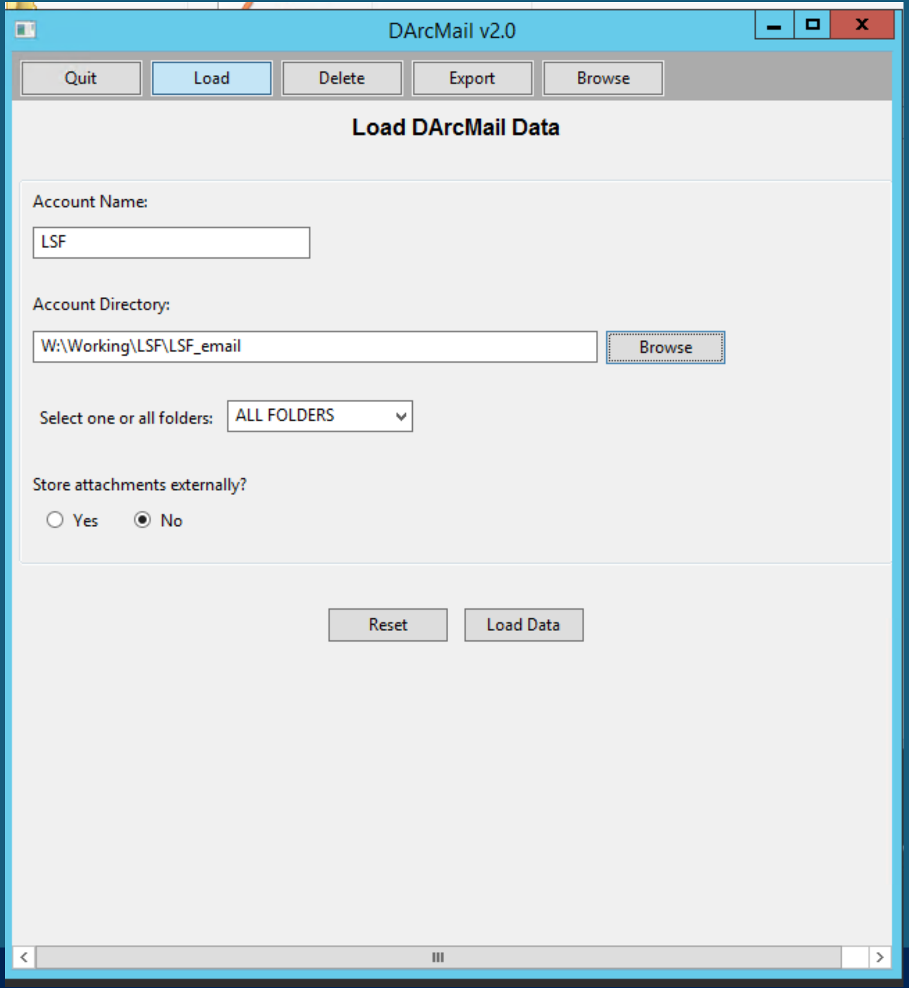
```
########## SETTINGS ##########
account: lsf3
account directory: w:/working/lsf/lsf3
external content: all attachments
chunk size: approximately 1073641824 bytes
output file: w:/working/lsf/lsf3\darcmail_test.xml

########## WARNINGS ##########
skipping duplicate message: <cm.0508451198012.butjykl.qdldjlidd.t@cmail20.com>
in folder darcmail_test (1 times)

########## SUMMARY ##########
total messages in mbox file(s): 6
duplicate messages skipped: 1
total messages in XML output: 5
external content files: 1
```

DArcMail functions



Connect to DArcMail Database

Sqlite3 database file: W:\Working\Tools\Software\DArcMail\2021_john_doe



Load DArcMail Data

Account Name: LSF

Account Directory: W:\Working\LSF\LSF_email

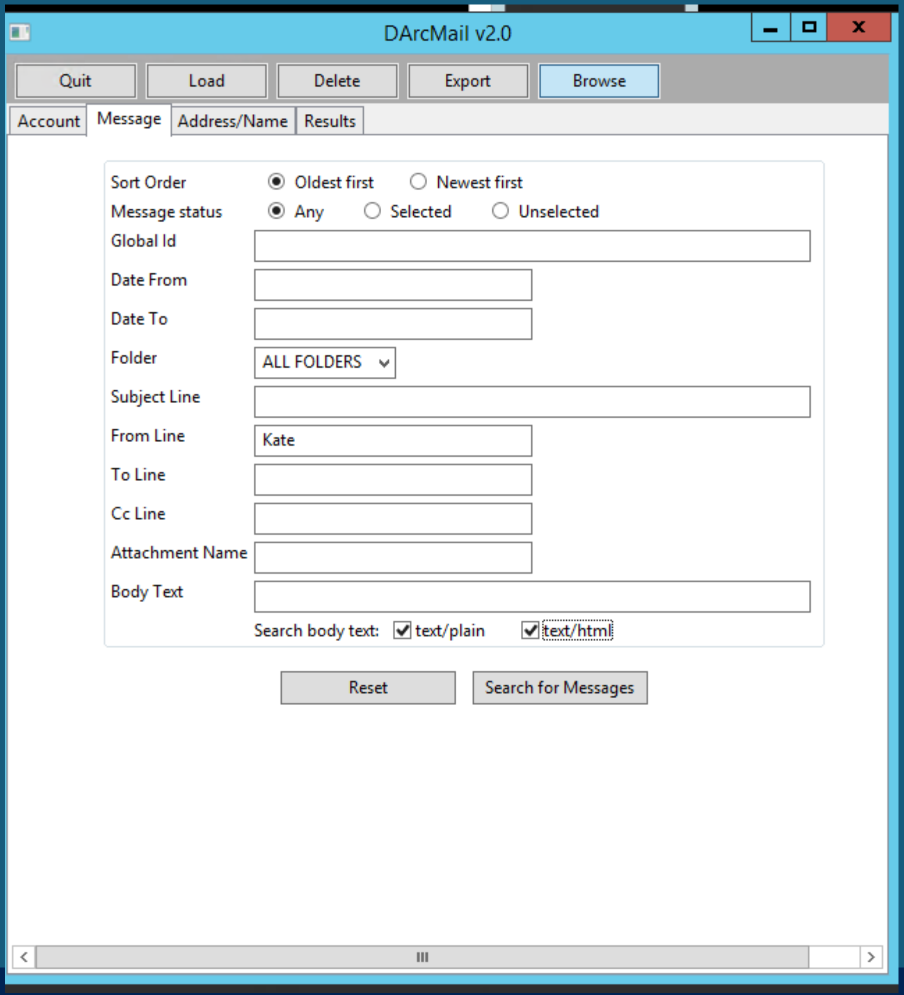
Select one or all folders: ALL FOLDERS

Store attachments externally?

○ Yes ● No

Smithsonian
Libraries and Archives

Review, Access - DArcMail



Sort Order ● Oldest first ○ Newest first

Message status ● Any ○ Selected ○ Unselected

Global Id

Date From

Date To

Folder ALL FOLDERS

Subject Line

From Line Kate

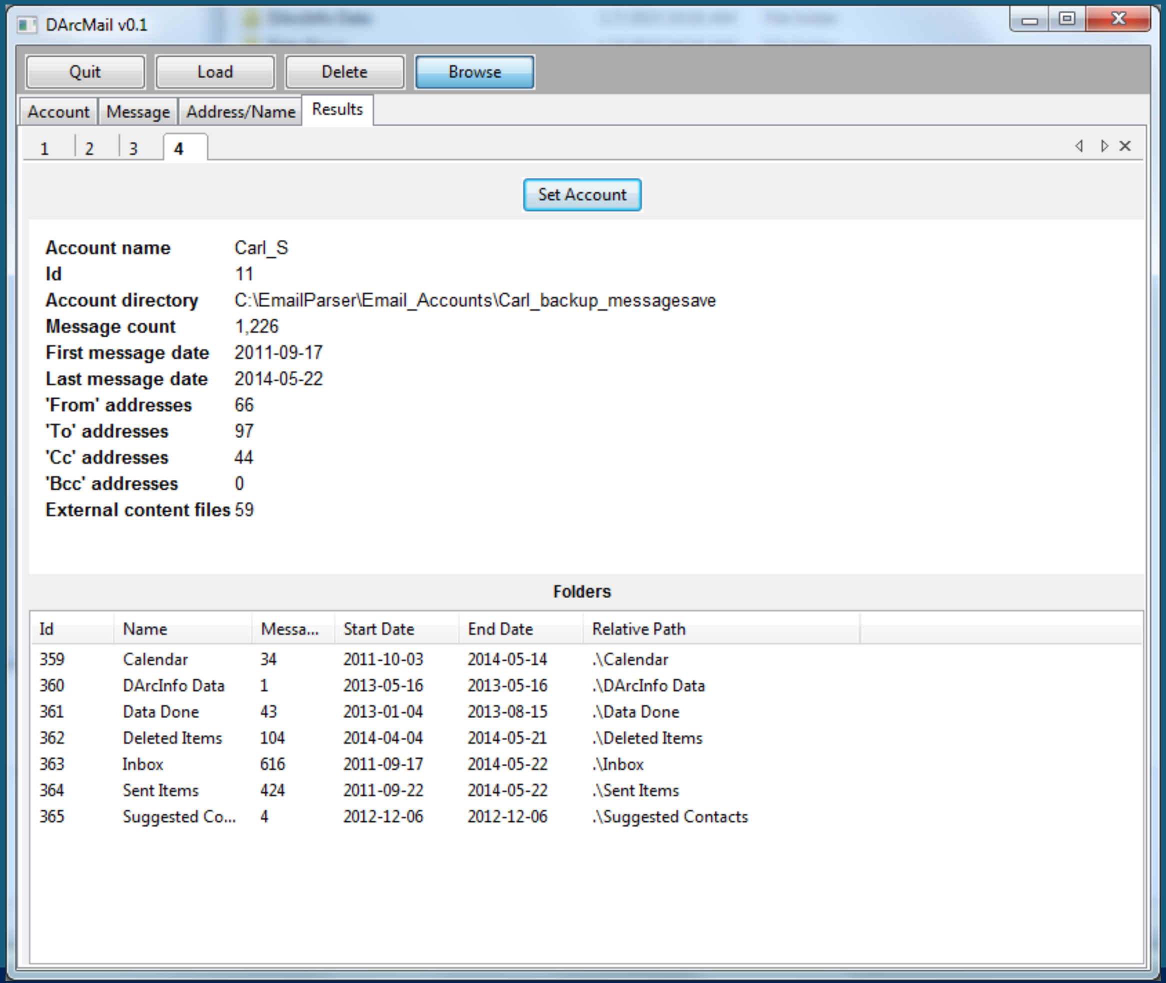
To Line

Cc Line

Attachment Name

Body Text

Search body text: ☑ text/plain ☑ text/html



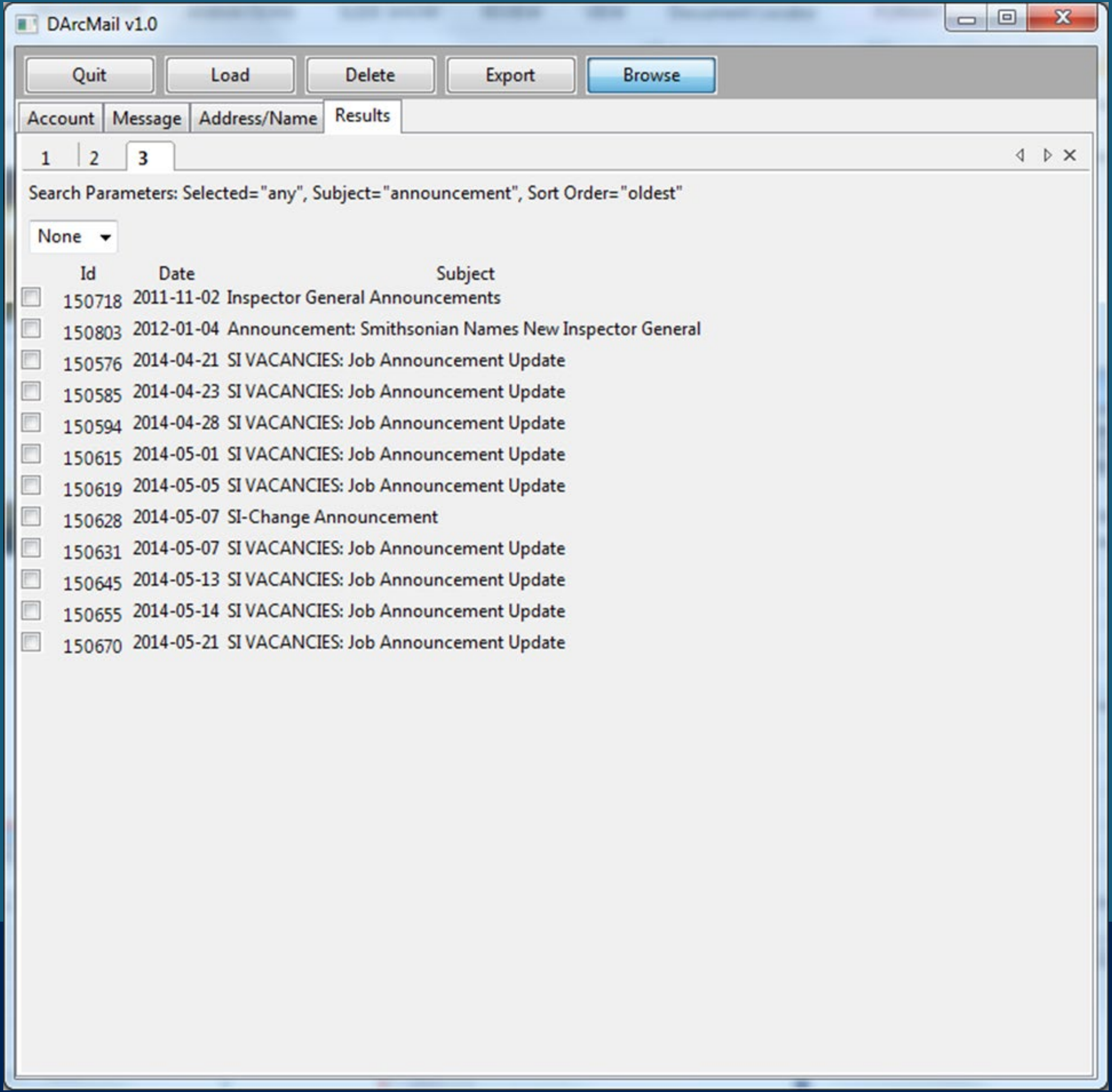
Account name	Carl_S
Id	11
Account directory	C:\EmailParser\Email_Accounts\Carl_backup_messagesave
Message count	1,226
First message date	2011-09-17
Last message date	2014-05-22
'From' addresses	66
'To' addresses	97
'Cc' addresses	44
'Bcc' addresses	0
External content files	59

Folders

Id	Name	Messa...	Start Date	End Date	Relative Path
359	Calendar	34	2011-10-03	2014-05-14	.\Calendar
360	DArcInfo Data	1	2013-05-16	2013-05-16	.\DArcInfo Data
361	Data Done	43	2013-01-04	2013-08-15	.\Data Done
362	Deleted Items	104	2014-04-04	2014-05-21	.\Deleted Items
363	Inbox	616	2011-09-17	2014-05-22	.\Inbox
364	Sent Items	424	2011-09-22	2014-05-22	.\Sent Items
365	Suggested Co...	4	2012-12-06	2012-12-06	.\Suggested Contacts

Smithsonian
Libraries and Archives

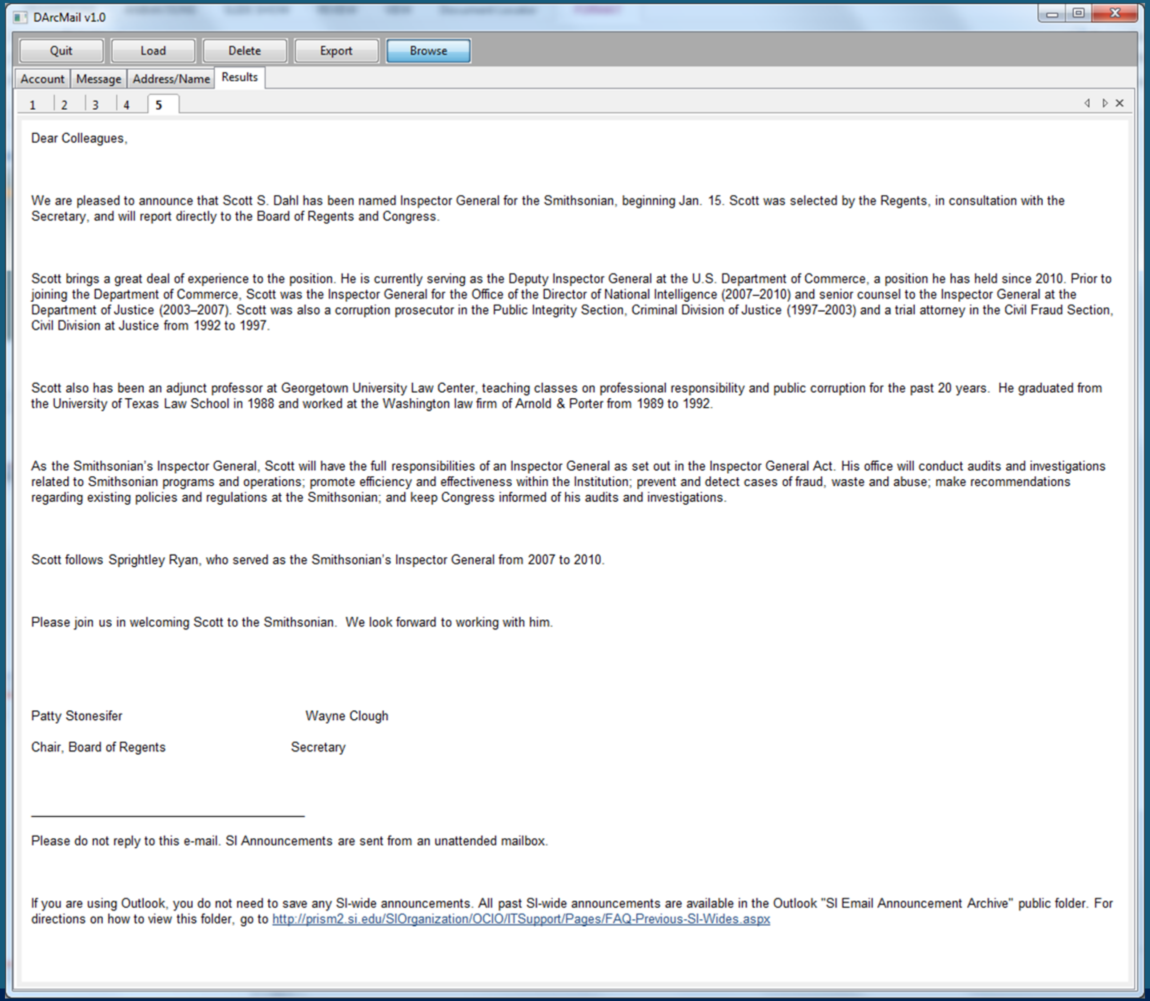
Search results



Search Parameters: Selected="any", Subject="announcement", Sort Order="oldest"

	Id	Date	Subject
□	150718	2011-11-02	Inspector General Announcements
□	150803	2012-01-04	Announcement: Smithsonian Names New Inspector General
□	150576	2014-04-21	SI VACANCIES: Job Announcement Update
□	150585	2014-04-23	SI VACANCIES: Job Announcement Update
□	150594	2014-04-28	SI VACANCIES: Job Announcement Update
□	150615	2014-05-01	SI VACANCIES: Job Announcement Update
□	150619	2014-05-05	SI VACANCIES: Job Announcement Update
□	150628	2014-05-07	SI-Change Announcement
□	150631	2014-05-07	SI VACANCIES: Job Announcement Update
□	150645	2014-05-13	SI VACANCIES: Job Announcement Update
□	150655	2014-05-14	SI VACANCIES: Job Announcement Update
□	150670	2014-05-21	SI VACANCIES: Job Announcement Update

Message view



Dear Colleagues,

We are pleased to announce that Scott S. Dahl has been named Inspector General for the Smithsonian, beginning Jan. 15. Scott was selected by the Regents, in consultation with the Secretary, and will report directly to the Board of Regents and Congress.

Scott brings a great deal of experience to the position. He is currently serving as the Deputy Inspector General at the U.S. Department of Commerce, a position he has held since 2010. Prior to joining the Department of Commerce, Scott was the Inspector General for the Office of the Director of National Intelligence (2007–2010) and senior counsel to the Inspector General at the Department of Justice (2003–2007). Scott was also a corruption prosecutor in the Public Integrity Section, Criminal Division of Justice (1997–2003) and a trial attorney in the Civil Fraud Section, Civil Division at Justice from 1992 to 1997.

Scott also has been an adjunct professor at Georgetown University Law Center, teaching classes on professional responsibility and public corruption for the past 20 years. He graduated from the University of Texas Law School in 1988 and worked at the Washington law firm of Arnold & Porter from 1989 to 1992.

As the Smithsonian's Inspector General, Scott will have the full responsibilities of an Inspector General as set out in the Inspector General Act. His office will conduct audits and investigations related to Smithsonian programs and operations; promote efficiency and effectiveness within the Institution; prevent and detect cases of fraud, waste and abuse; make recommendations regarding existing policies and regulations at the Smithsonian; and keep Congress informed of his audits and investigations.

Scott follows Sprightley Ryan, who served as the Smithsonian's Inspector General from 2007 to 2010.

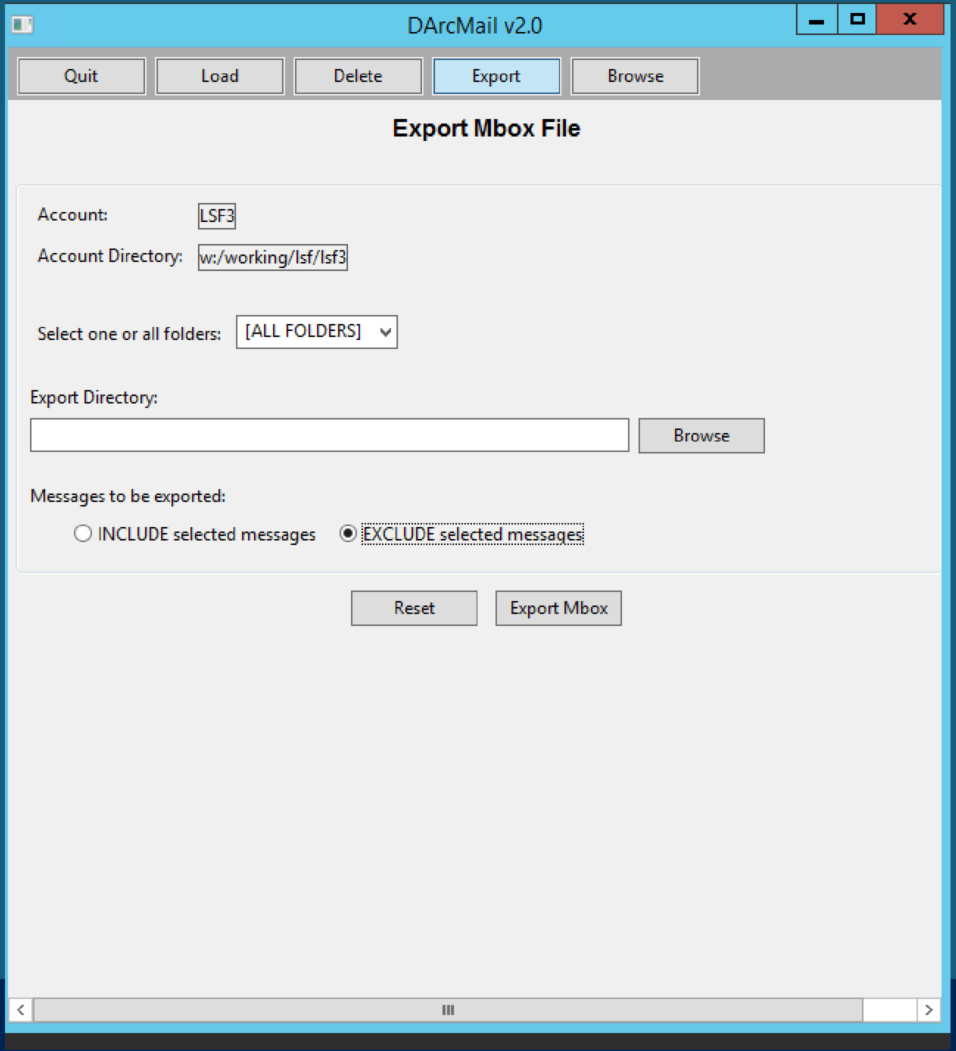
Please join us in welcoming Scott to the Smithsonian. We look forward to working with him.

Patty Stonesifer Wayne Clough
Chair, Board of Regents Secretary

Please do not reply to this e-mail. SI Announcements are sent from an unattended mailbox.

If you are using Outlook, you do not need to save any SI-wide announcements. All past SI-wide announcements are available in the Outlook "SI Email Announcement Archive" public folder. For directions on how to view this folder, go to http://prism2.si.edu/SIOrganization/OCIO/ITSupport/Pages/FAQ-Previous-SI-Wides.aspx

Export selection to new MBOX file



Export Mbox File

Account: LSF3

Account Directory: w:/working/lsf/lsf3

Select one or all folders: [ALL FOLDERS]

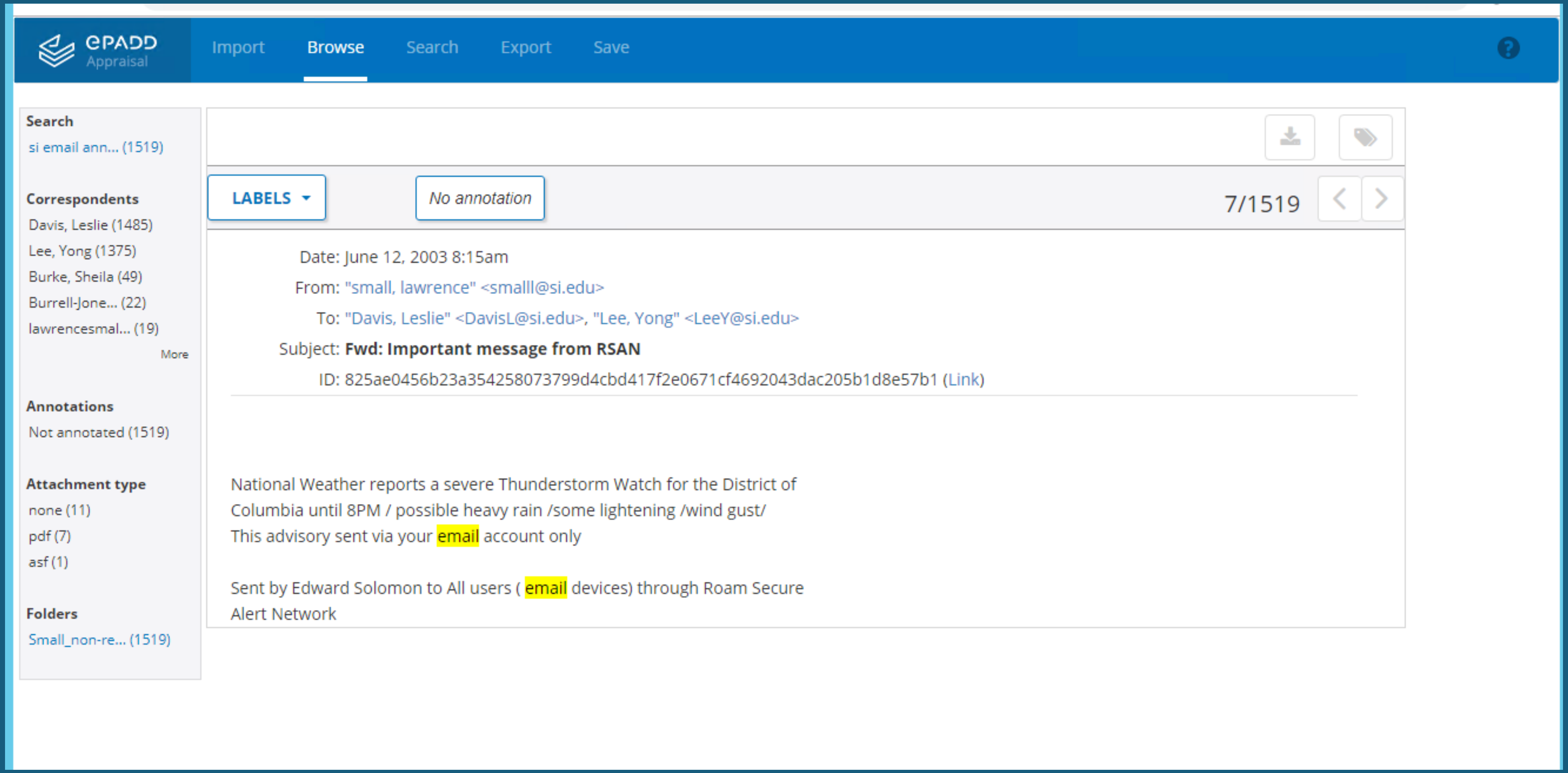
Export Directory:

Messages to be exported:

○ INCLUDE selected messages ● EXCLUDE selected messages

Smithsonian
Libraries and Archives

Review, Access - ePADD



Search

si email ann... (1519)

Correspondents

Davis, Leslie (1485)
Lee, Yong (1375)
Burke, Sheila (49)
Burrell-Jone... (22)
lawrencesmal... (19)
More

Annotations

Not annotated (1519)

Attachment type

none (11)
pdf (7)
asf (1)

Folders

Small_non-re... (1519)

Date: June 12, 2003 8:15am
From: "small, lawrence" <smalll@si.edu>
To: "Davis, Leslie" <DavisL@si.edu>, "Lee, Yong" <LeeY@si.edu>
Subject: **Fwd: Important message from RSAN**
ID: 825ae0456b23a354258073799d4cbd417f2e0671cf4692043dac205b1d8e57b1 (Link)

National Weather reports a severe Thunderstorm Watch for the District of Columbia until 8PM / possible heavy rain /some lightening /wind gust/
This advisory sent via your email account only

Sent by Edward Solomon to All users (email devices) through Roam Secure Alert Network

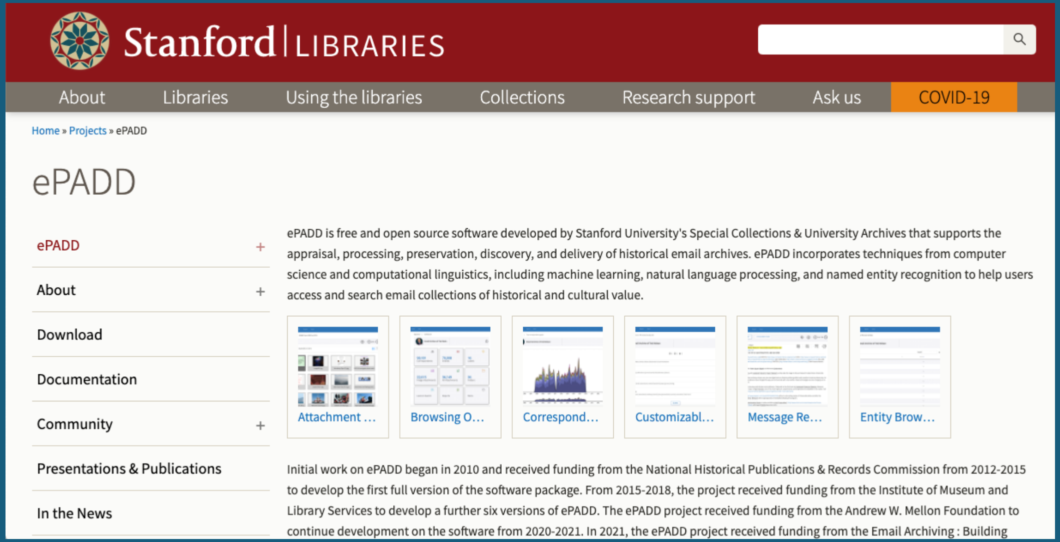
2020s

- Larger accounts
- More people leaving
  = more accessions
- Virus scan tool discontinued

National Postal Museum,
Accession Number: A.2006-34

Smithsonian
Libraries and Archives

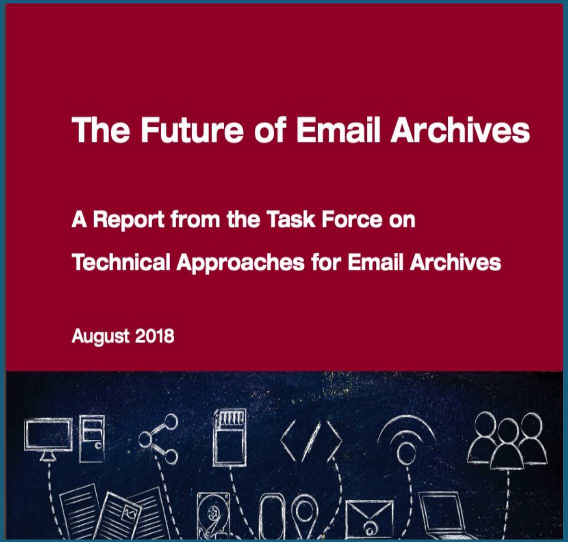
Email Projects/Events/Reports



ePADD

ePADD is free and open source software developed by Stanford University's Special Collections & University Archives that supports the appraisal, processing, preservation, discovery, and delivery of historical email archives. ePADD incorporates techniques from computer science and computational linguistics, including machine learning, natural language processing, and named entity recognition to help users access and search email collections of historical and cultural value.

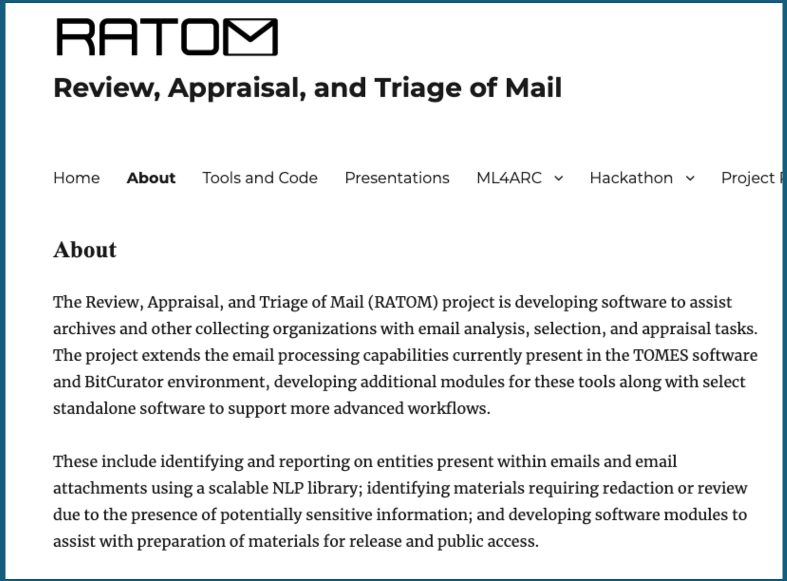
Initial work on ePADD began in 2010 and received funding from the National Historical Publications & Records Commission from 2012-2015 to develop the first full version of the software package. From 2015-2018, the project received funding from the Institute of Museum and Library Services to develop a further six versions of ePADD. The ePADD project received funding from the Andrew W. Mellon Foundation to continue development on the software from 2020-2021. In 2021, the ePADD project received funding from the Email Archiving : Building



The Future of Email Archives

A Report from the Task Force on Technical Approaches for Email Archives

August 2018



RATOM

Review, Appraisal, and Triage of Mail

About

The Review, Appraisal, and Triage of Mail (RATOM) project is developing software to assist archives and other collecting organizations with email analysis, selection, and appraisal tasks. The project extends the email processing capabilities currently present in the TOMES software and BitCurator environment, developing additional modules for these tools along with select standalone software to support more advanced workflows.

These include identifying and reporting on entities present within emails and email attachments using a scalable NLP library; identifying materials requiring redaction or review due to the presence of potentially sensitive information; and developing software modules to assist with preparation of materials for release and public access.

LIBRARY LIBRARY OF CONGRESS
Everything
Search Loc.gov
Share
FILM, VIDEO

Archiving Email: Institutional Approaches to Processing & Archiving Email

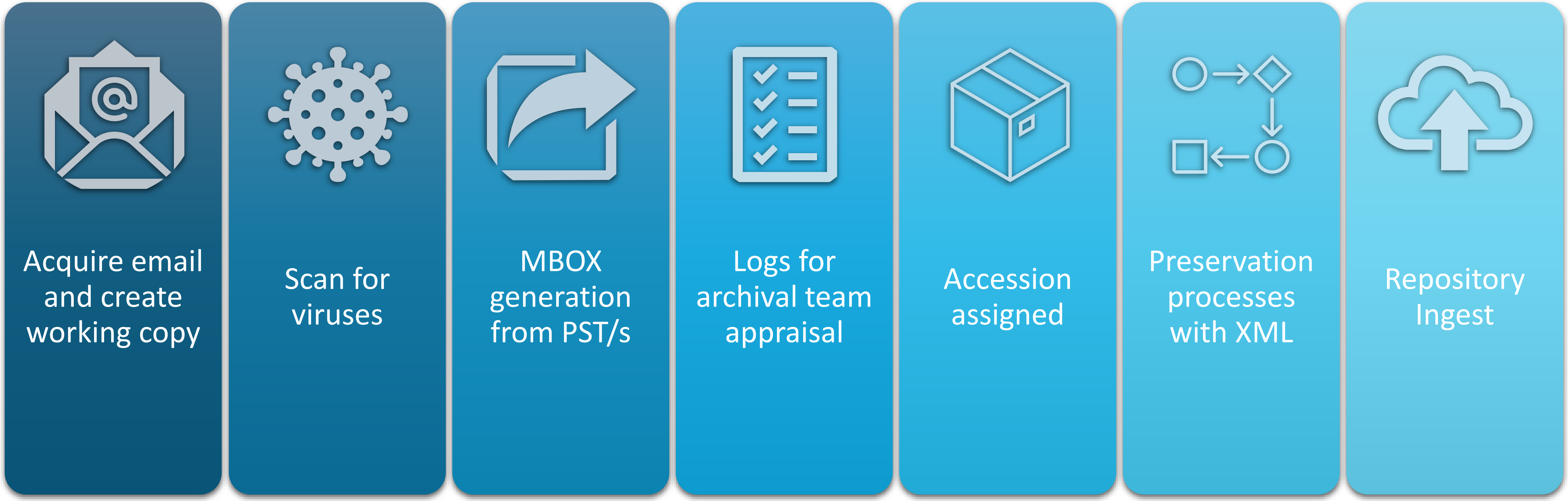
CO-HOSTED BY THE LIBRARY OF CONGRESS AND THE NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
Archiving Email Symposium: Technical Perspective: Institutional Approaches to Processing and Archiving Email
Various Speakers
June 2, 2015

A SPECIFICATION FOR USING PDF TO PACKAGE AND REPRESENT EMAIL

EA-PDF WORKING GROUP

TECHNICAL REPORT PUBLISHED BY
THE UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Previous workflow



Acquire email and create working copy

Scan for viruses

MBOX generation from PST/s

Logs for archival team appraisal

Accession assigned

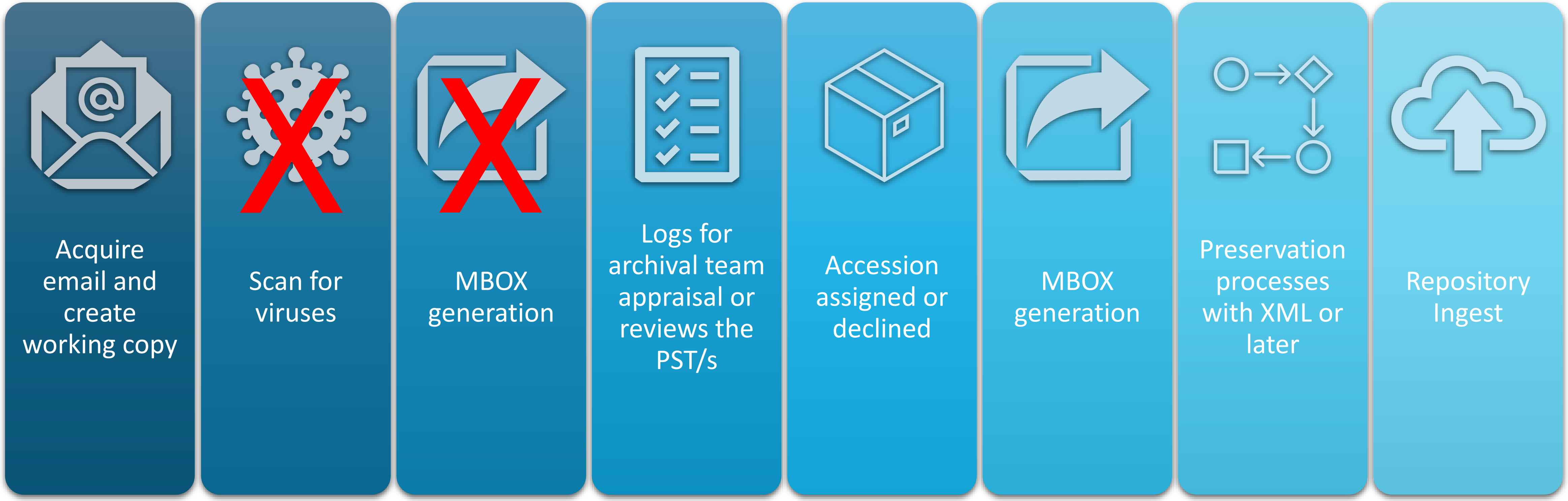
Preservation processes with XML

Repository Ingest

Bottleneck here due to time to take to complete

Smithsonian
Libraries and Archives

Current workflow



Acquire email and create working copy

Scan for viruses

MBOX generation

Logs for archival team appraisal or reviews the PST/s

Accession assigned or declined

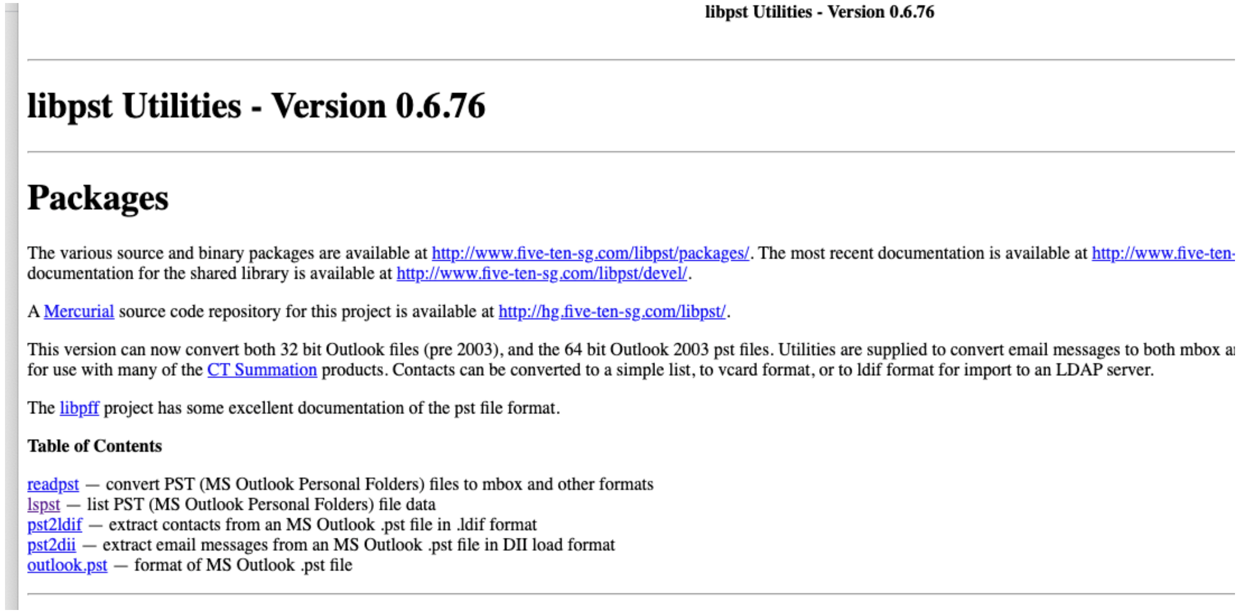
MBOX generation

Preservation processes with XML or later

Repository Ingest

Smithsonian
Libraries and Archives

Other tools



libpst Utilities - Version 0.6.76

libpst Utilities - Version 0.6.76

Packages

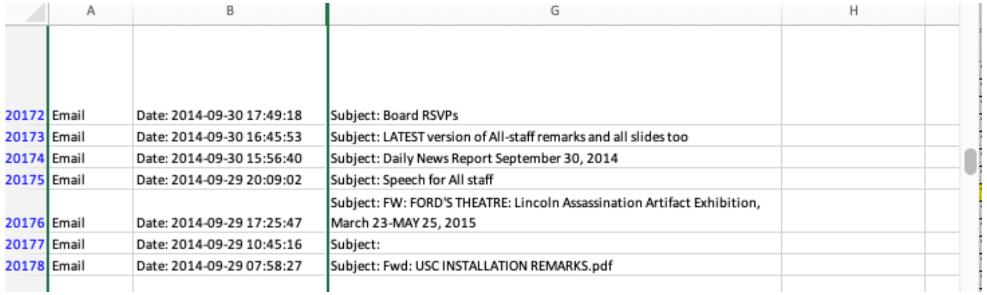
The various source and binary packages are available at http://www.five-ten-sg.com/libpst/packages/. The most recent documentation is available at http://www.five-ten-
documentation for the shared library is available at http://www.five-ten-sg.com/libpst/devel/.

A Mercurial source code repository for this project is available at http://hg.five-ten-sg.com/libpst/.

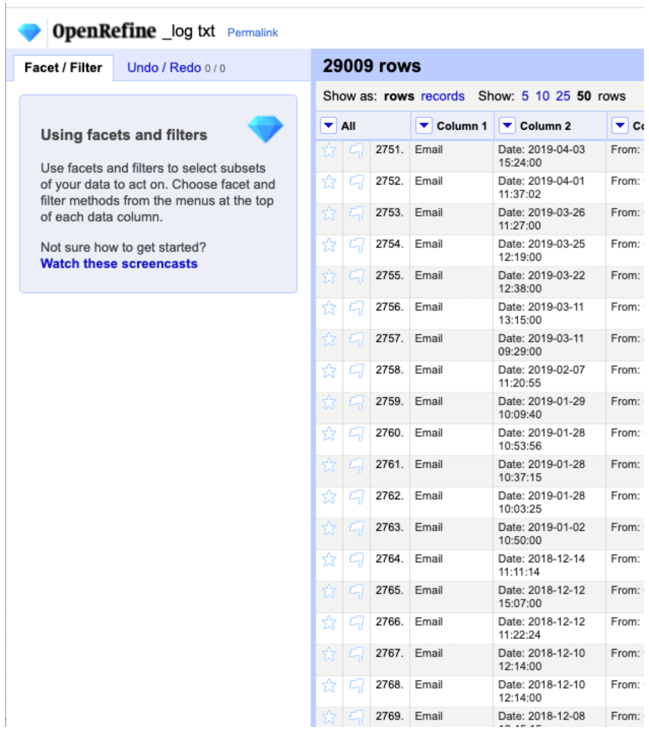
This version can now convert both 32 bit Outlook files (pre 2003), and the 64 bit Outlook 2003 pst files. Utilities are supplied to convert email messages to both mbox a
for use with many of the CT Summation products. Contacts can be converted to a simple list, to vcard format, or to ldif format for import to an LDAP server.

The libpff project has some excellent documentation of the pst file format.

Table of Contents



	A	B	G	H
20172	Email	Date: 2014-09-30 17:49:18	Subject: Board RSVPs	
20173	Email	Date: 2014-09-30 16:45:53	Subject: LATEST version of All-staff remarks and all slides too	
20174	Email	Date: 2014-09-30 15:56:40	Subject: Daily News Report September 30, 2014	
20175	Email	Date: 2014-09-29 20:09:02	Subject: Speech for All staff	
20176	Email	Date: 2014-09-29 17:25:47	Subject: FW: FORD'S THEATRE: Lincoln Assassination Artifact Exhibition, March 23-MAY 25, 2015	
20177	Email	Date: 2014-09-29 10:45:16	Subject:	
20178	Email	Date: 2014-09-29 07:58:27	Subject: Fwd: USC INSTALLATION REMARKS.pdf	



Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

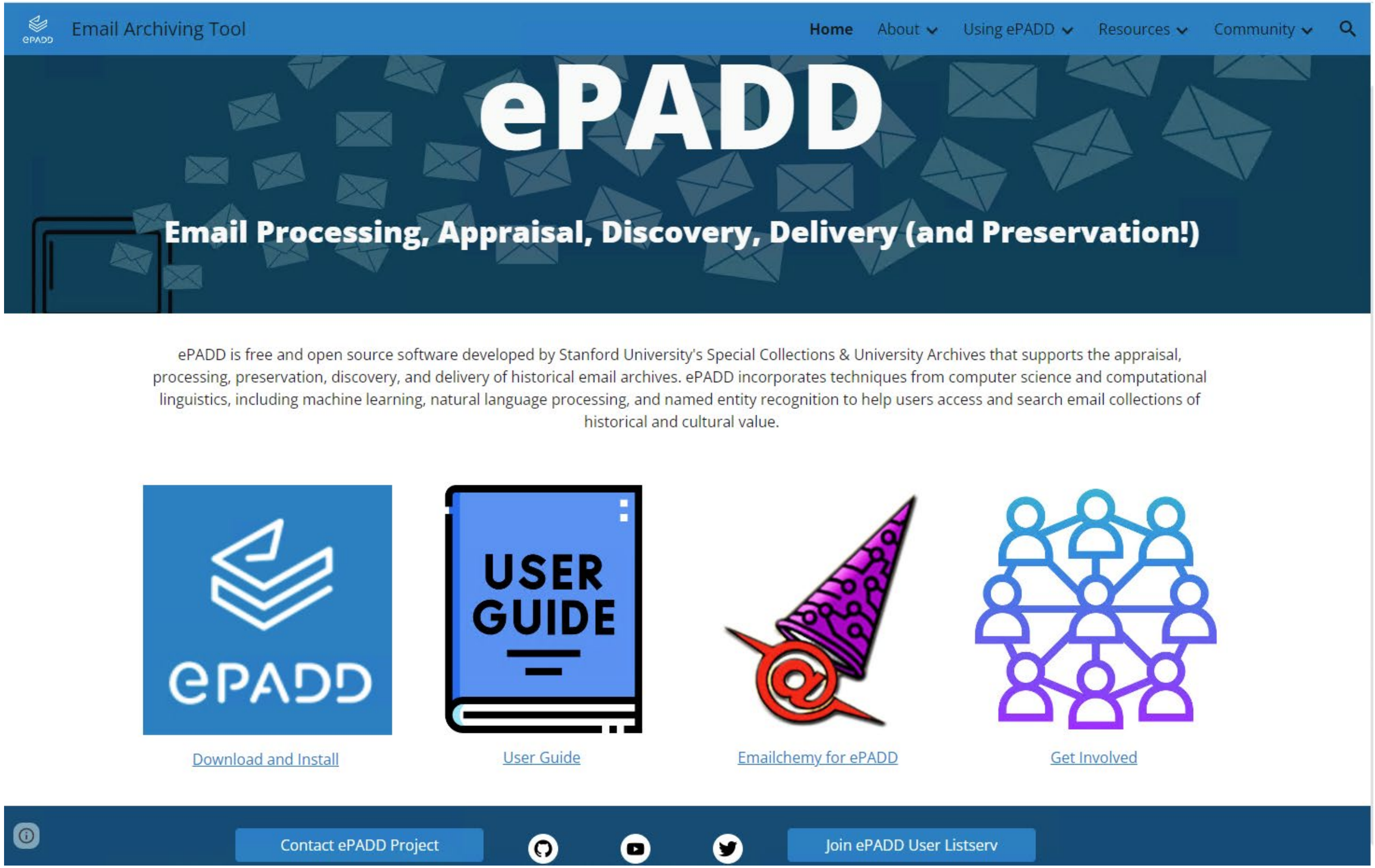
29009 rows

Show as: **rows** records Show: 5 10 25 **50** rows

All	Column 1	Column 2	Co
2751.	Email	Date: 2019-04-03 15:24:00	From:
2752.	Email	Date: 2019-04-01 11:37:02	From:
2753.	Email	Date: 2019-03-26 11:27:00	From:
2754.	Email	Date: 2019-03-25 12:19:00	From:
2755.	Email	Date: 2019-03-22 12:38:00	From:
2756.	Email	Date: 2019-03-11 13:15:00	From:
2757.	Email	Date: 2019-03-11 09:29:00	From:
2758.	Email	Date: 2019-02-07 11:20:55	From:
2759.	Email	Date: 2019-01-29 10:09:40	From:
2760.	Email	Date: 2019-01-28 10:53:56	From:
2761.	Email	Date: 2019-01-28 10:37:15	From:
2762.	Email	Date: 2019-01-28 10:03:25	From:
2763.	Email	Date: 2019-01-02 10:50:00	From:
2764.	Email	Date: 2018-12-14 11:11:14	From:
2765.	Email	Date: 2018-12-12 15:07:00	From:
2766.	Email	Date: 2018-12-12 11:22:24	From:
2767.	Email	Date: 2018-12-10 12:14:00	From:
2768.	Email	Date: 2018-12-10 12:14:00	From:
2769.	Email	Date: 2018-12-08	From:

Smithsonian
Libraries and Archives

ePADD today – Now accepting PSTs!

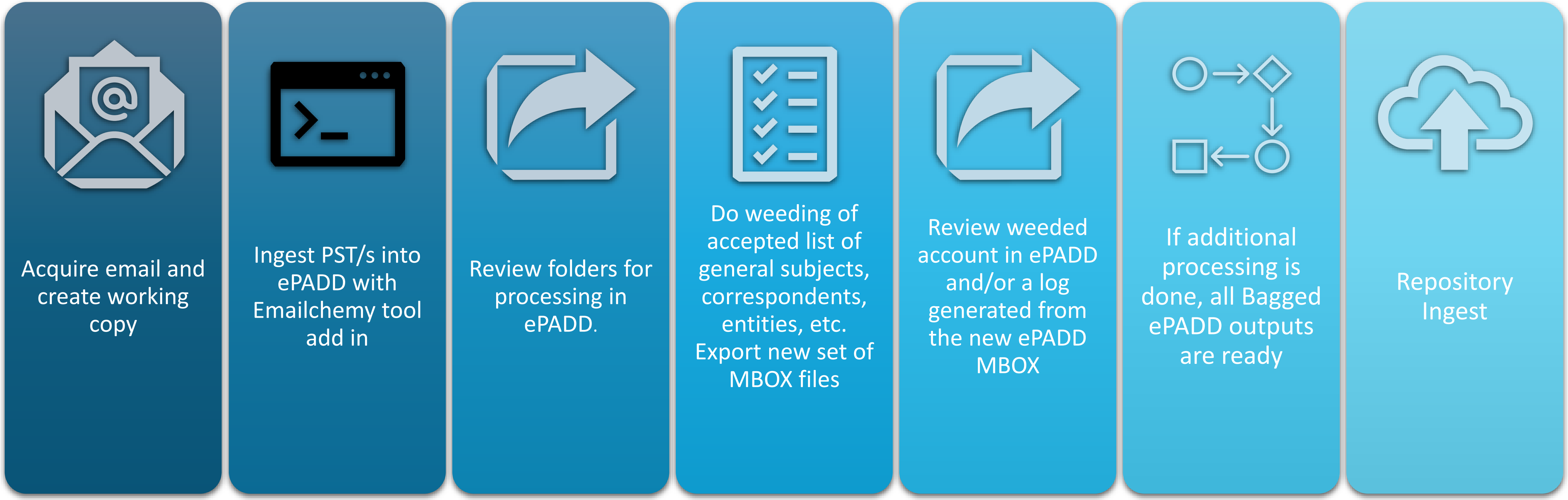


Email Archiving Tool
Home
About
Using ePADD
Resources
Community
ePADD
Email Processing, Appraisal, Discovery, Delivery (and Preservation!)

ePADD is free and open source software developed by Stanford University's Special Collections & University Archives that supports the appraisal, processing, preservation, discovery, and delivery of historical email archives. ePADD incorporates techniques from computer science and computational linguistics, including machine learning, natural language processing, and named entity recognition to help users access and search email collections of historical and cultural value.

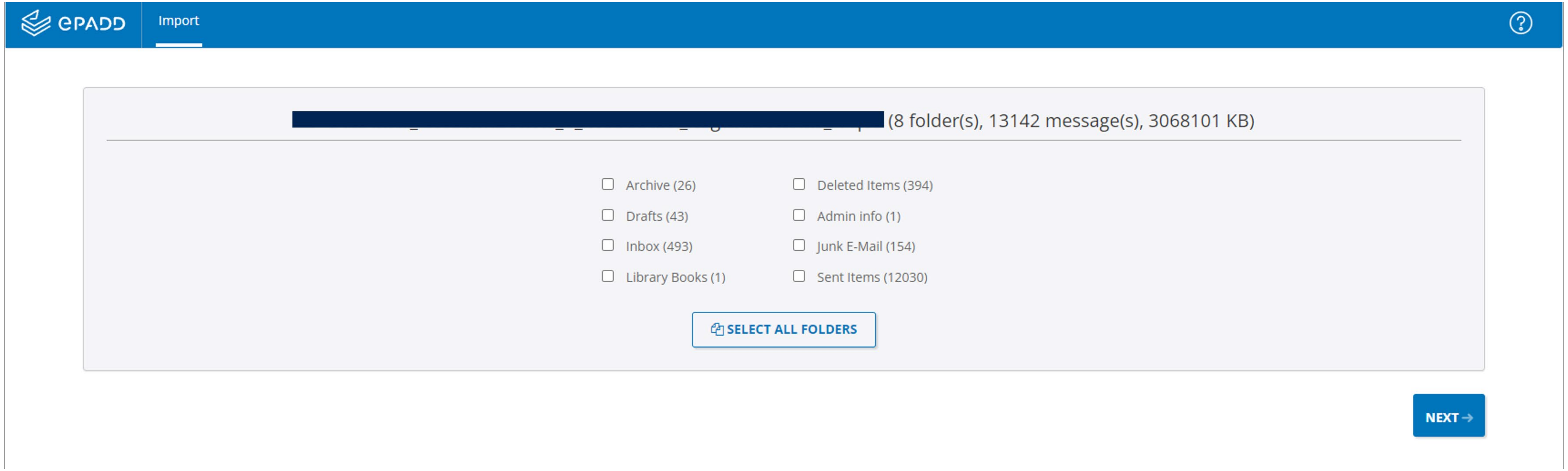
Smithsonian
Libraries and Archives

Potential workflow



1. Acquire email and create working copy
2. Ingest PST/s into ePADD with Emailchemy tool add in
3. Review folders for processing in ePADD.
4. Do weeding of accepted list of general subjects, correspondents, entities, etc. Export new set of MBOX files
5. Review weeded account in ePADD and/or a log generated from the new ePADD MBOX
6. If additional processing is done, all Bagged ePADD outputs are ready
7. Repository Ingest

Smithsonian
Libraries and Archives



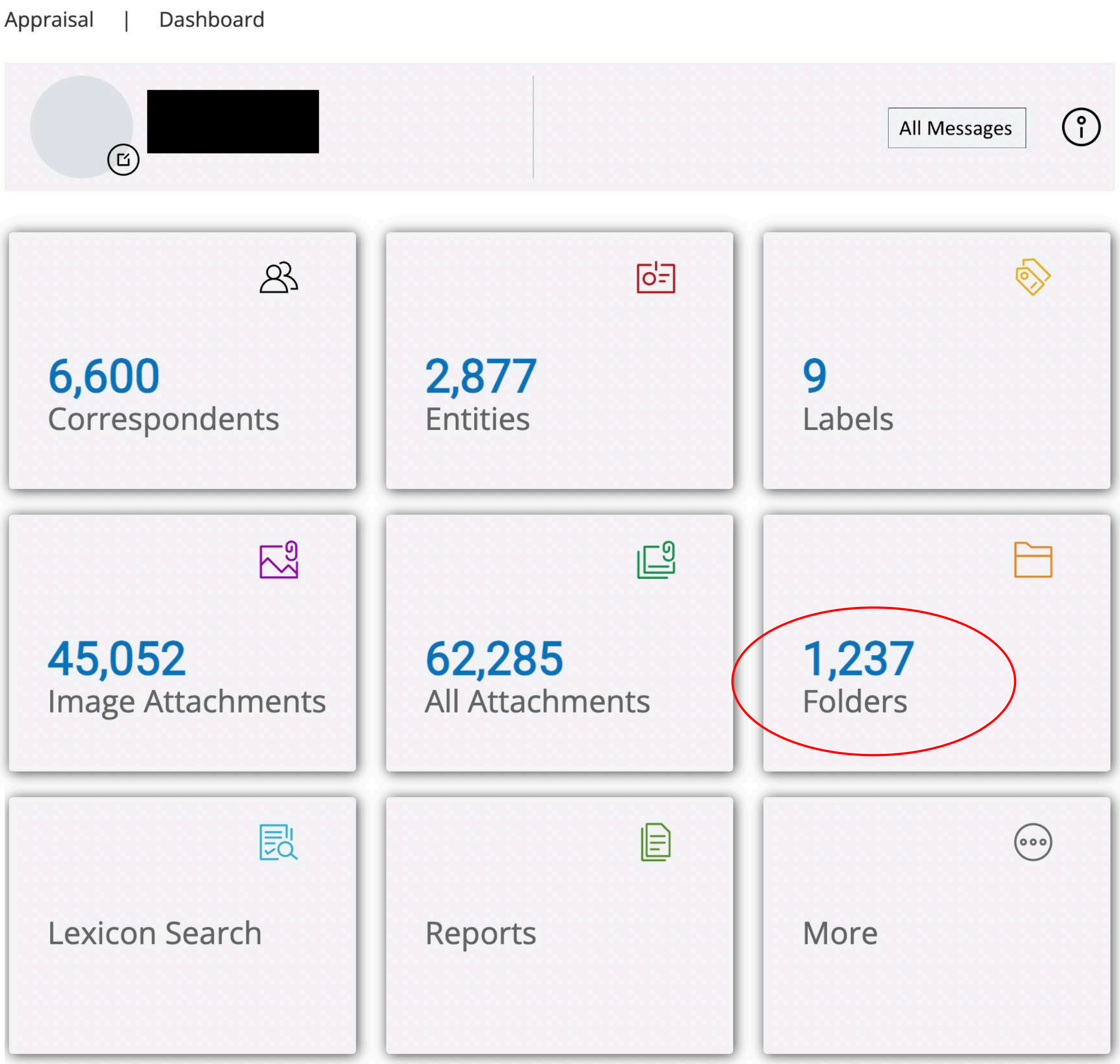
[redacted] (8 folder(s), 13142 message(s), 3068101 KB)

□ Archive (26)
□ Drafts (43)
□ Inbox (493)
□ Library Books (1)
□ Deleted Items (394)
□ Admin info (1)
□ Junk E-Mail (154)
□ Sent Items (12030)

- **Don’t transfer directories**
  - **Junk**
  - **Personal**
  - **Personnel**
  - **Deleted Items – sometimes ***

11.5 GB account

More than 53,000 messages



Appraisal | Dashboard

[redacted]

6,600 Correspondents

2,877 Entities

9 Labels

45,052 Image Attachments

62,285 All Attachments

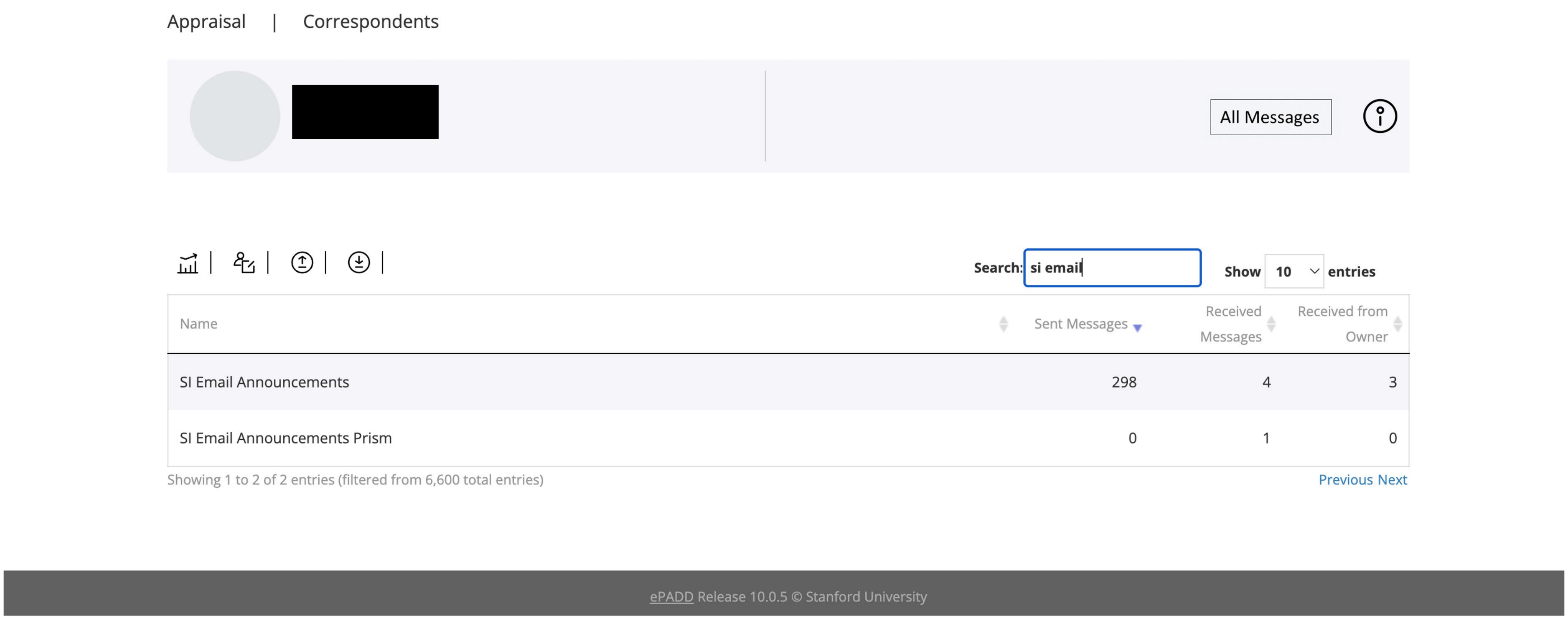
1,237 Folders

Lexicon Search

Reports

More

Weeding – do not transfer



[redacted]

Search: si email

Show 10 entries

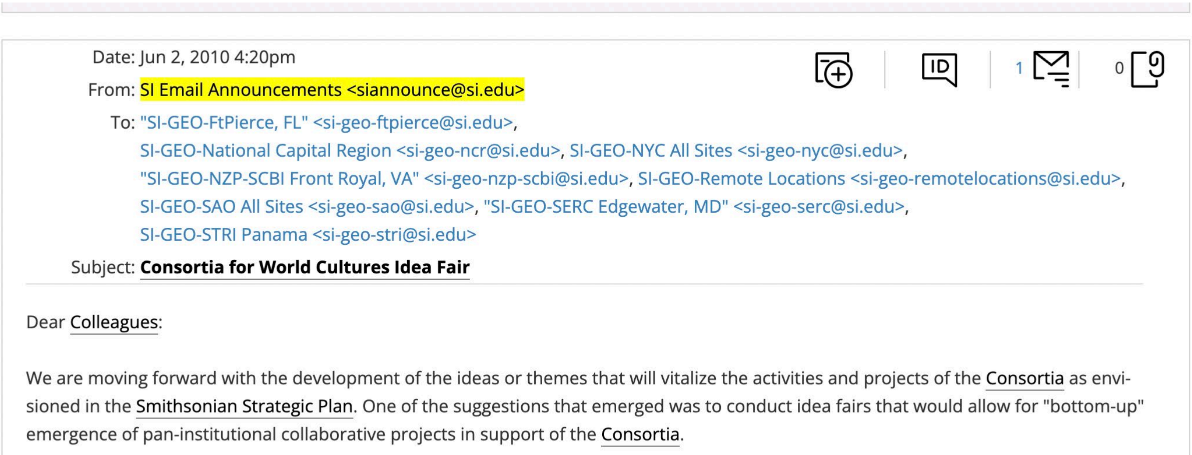
Name	Sent Messages	Received Messages	Received from Owner
SI Email Announcements	298	4	3
SI Email Announcements Prism	0	1	0

Showing 1 to 2 of 2 entries (filtered from 6,600 total entries)

ePADD Release 10.0.5 © Stanford University

Smithsonian
Libraries and Archives

Selecting Items



Date: Jun 2, 2010 4:20pm

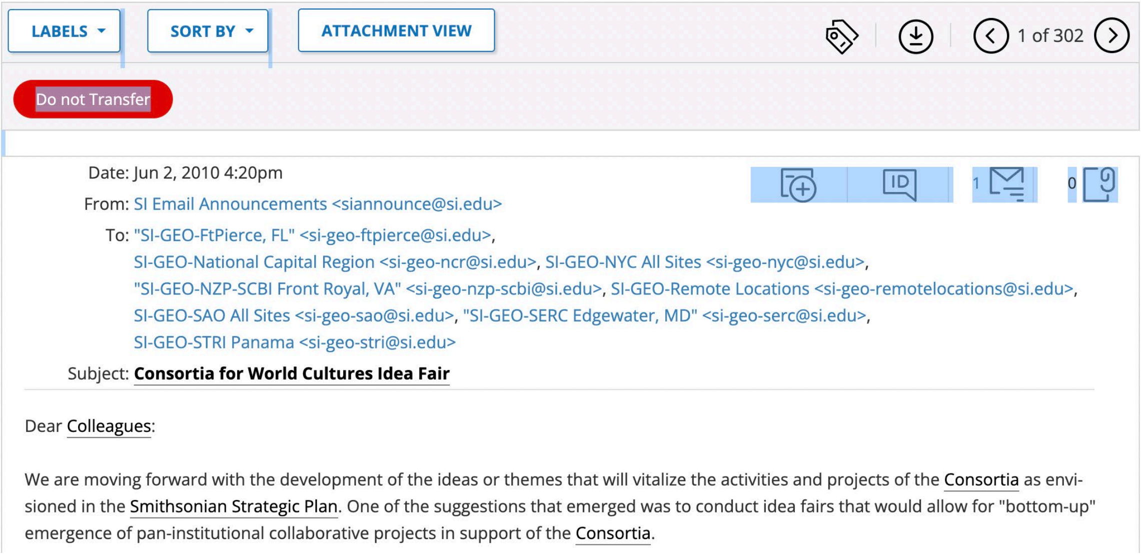
From: SI Email Announcements <siannounce@si.edu>

To: "SI-GEO-FtPierce, FL" <si-geo-ftpierce@si.edu>, SI-GEO-National Capital Region <si-geo-ncr@si.edu>, SI-GEO-NYC All Sites <si-geo-nyc@si.edu>, "SI-GEO-NZP-SCBI Front Royal, VA" <si-geo-nzp-scbi@si.edu>, SI-GEO-Remote Locations <si-geo-remotelocations@si.edu>, SI-GEO-SAO All Sites <si-geo-sao@si.edu>, "SI-GEO-SERC Edgewater, MD" <si-geo-serc@si.edu>, SI-GEO-STRI Panama <si-geo-stri@si.edu>

Subject: **Consortia for World Cultures Idea Fair**

Dear Colleagues:

We are moving forward with the development of the ideas or themes that will vitalize the activities and projects of the Consortia as envisioned in the Smithsonian Strategic Plan. One of the suggestions that emerged was to conduct idea fairs that would allow for "bottom-up" emergence of pan-institutional collaborative projects in support of the Consortia.



Do not Transfer

Date: Jun 2, 2010 4:20pm

From: SI Email Announcements <siannounce@si.edu>

To: "SI-GEO-FtPierce, FL" <si-geo-ftpierce@si.edu>, SI-GEO-National Capital Region <si-geo-ncr@si.edu>, SI-GEO-NYC All Sites <si-geo-nyc@si.edu>, "SI-GEO-NZP-SCBI Front Royal, VA" <si-geo-nzp-scbi@si.edu>, SI-GEO-Remote Locations <si-geo-remotelocations@si.edu>, SI-GEO-SAO All Sites <si-geo-sao@si.edu>, "SI-GEO-SERC Edgewater, MD" <si-geo-serc@si.edu>, SI-GEO-STRI Panama <si-geo-stri@si.edu>

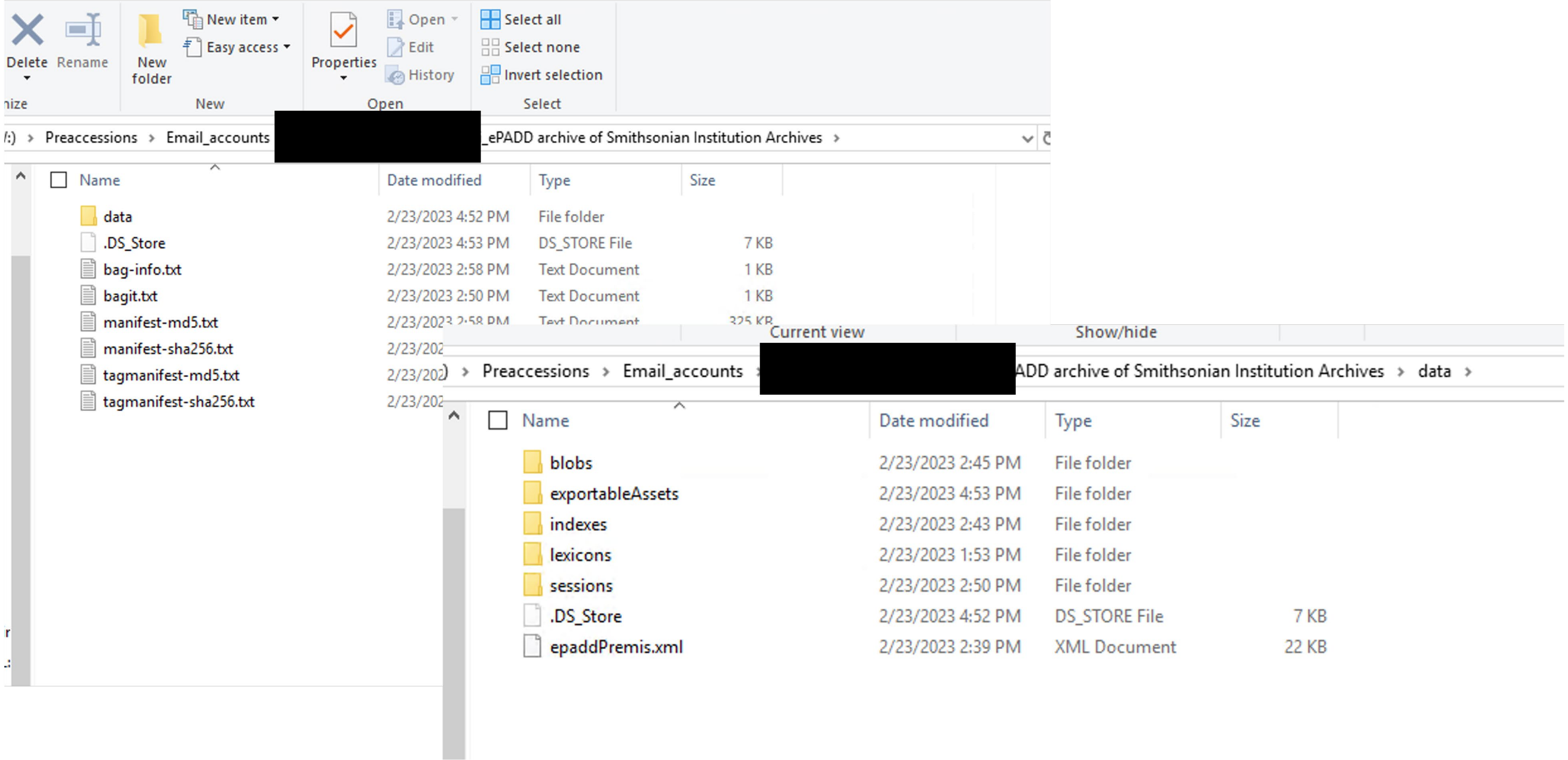
Subject: **Consortia for World Cultures Idea Fair**

Dear Colleagues:

We are moving forward with the development of the ideas or themes that will vitalize the activities and projects of the Consortia as envisioned in the Smithsonian Strategic Plan. One of the suggestions that emerged was to conduct idea fairs that would allow for "bottom-up" emergence of pan-institutional collaborative projects in support of the Consortia.

Smithsonian
Libraries and Archives

ePADD Bag



:) › Preaccessions › Email_accounts › [redacted] _ePADD archive of Smithsonian Institution Archives ›

Name	Date modified	Type	Size
data	2/23/2023 4:52 PM	File folder	
.DS_Store	2/23/2023 4:53 PM	DS_STORE File	7 KB
bag-info.txt	2/23/2023 2:58 PM	Text Document	1 KB
bagit.txt	2/23/2023 2:50 PM	Text Document	1 KB
manifest-md5.txt	2/23/2023 2:58 PM	Text Document	325 KB
manifest-sha256.txt	2/23/202		
tagmanifest-md5.txt	2/23/202		
tagmanifest-sha256.txt	2/23/202		

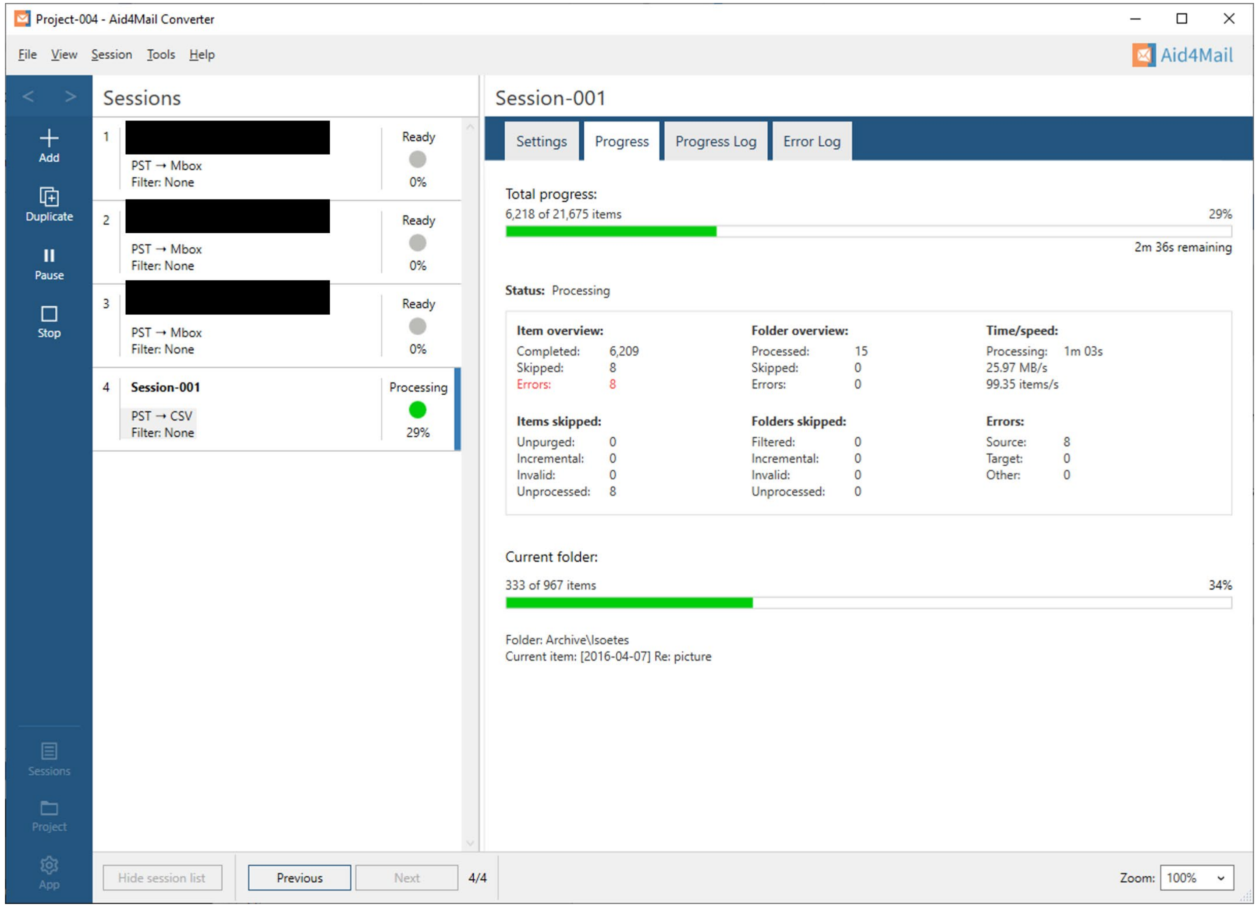
) › Preaccessions › Email_accounts › [redacted] ADD archive of Smithsonian Institution Archives › data ›

Name	Date modified	Type	Size
blobs	2/23/2023 2:45 PM	File folder	
exportableAssets	2/23/2023 4:53 PM	File folder	
indexes	2/23/2023 2:43 PM	File folder	
lexicons	2/23/2023 1:53 PM	File folder	
sessions	2/23/2023 2:50 PM	File folder	
.DS_Store	2/23/2023 4:52 PM	DS_STORE File	7 KB
epaddPremis.xml	2/23/2023 2:39 PM	XML Document	22 KB

Smithsonian
Libraries and Archives

First Impressions

- PSTs into ePADD – not complete. Calendar?
- CSVs don’t have subject line – but it’s coming
- Directory/subdirectory structure not retained on import into ePADD
- Orcids and Sharepoint URLs flagged as CCNs. Finding SSNs



Sessions

1 [redacted]
PST → Mbox
Filter: None
Ready
0%

2 [redacted]
PST → Mbox
Filter: None
Ready
0%

3 [redacted]
PST → Mbox
Filter: None
Ready
0%

4 Session-001
PST → CSV
Filter: None
Processing
29%

4/4

Session-001

Total progress:
6,218 of 21,675 items 29%
2m 36s remaining

Status: Processing

Item overview:		Folder overview:		Time/speed:	
Completed:	6,209	Processed:	15	Processing:	1m 03s
Skipped:	8	Skipped:	0	25.97 MB/s	
Errors:	8	Errors:	0	99.35 items/s	
Items skipped:		**Folders skipped:**		**Errors:**	
Unpurged:	0	Filtered:	0	Source:	8
Incremental:	0	Incremental:	0	Target:	0
Invalid:	0	Invalid:	0	Other:	0
Unprocessed:	8	Unprocessed:	0		

Current folder:
333 of 967 items 34%

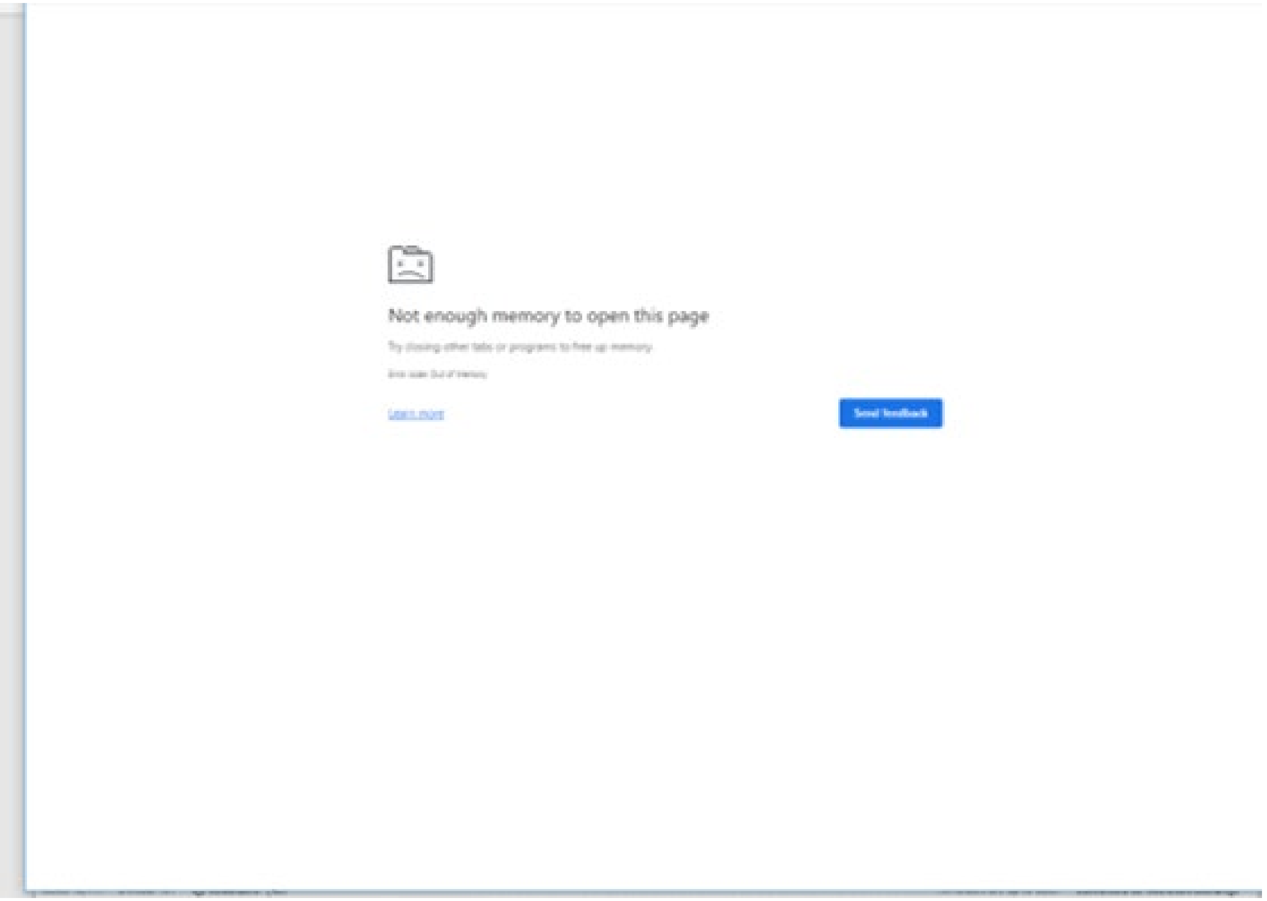
Folder: Archive\Isoetes
Current item: [2016-04-07] Re: picture

Zoom: 100%

Smithsonian
Libraries and Archives

First Impressions

- Accounts larger than 40 GB have crashed – revisit
- Some formatting issues in the messages. Nature of email though
- Faster way to review and “weed” for final version. ☺
- Appraising archivist – likes the ease of use and the authorities in processing module.



Not enough memory to open this page

Try closing other tabs or programs to free up memory

Error code: Out of memory

Moving ahead

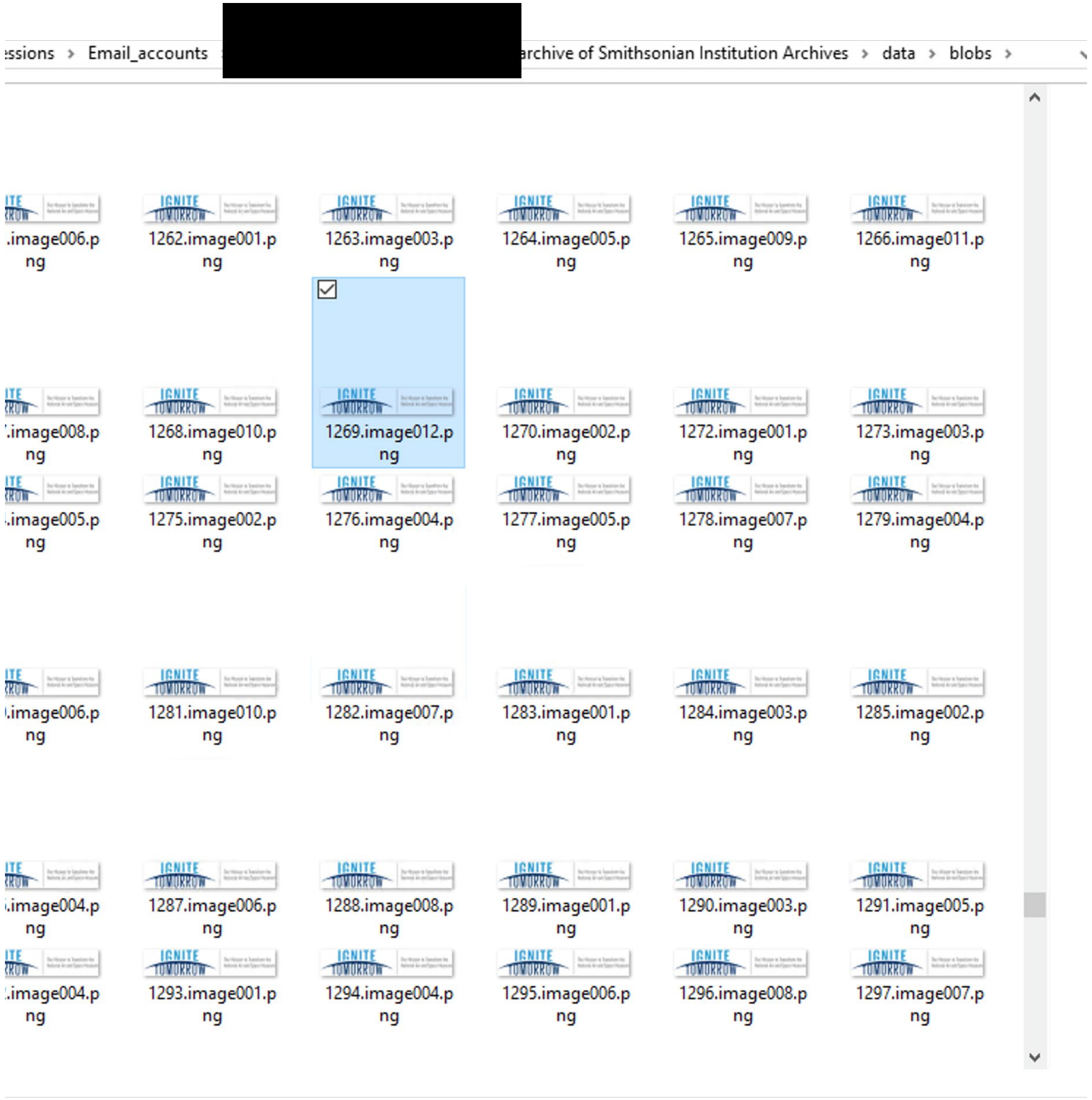
Where do the other previous tools fit in?

New release coming that will have the subject in the CSV. ☺

Tradeoff – willing to live without those Calendar items showing up for appraisal?

How will tasks and calendar items be preserved?

Appraisal – will the archivists want to learn ePADD for reviewing? Or back to the logs?



ssions › Email_accounts › [redacted] archive of Smithsonian Institution Archives › data › blobs ›

.image006.png 1262.image001.png 1263.image003.png 1264.image005.png 1265.image009.png 1266.image011.png

.image008.png 1268.image010.png ☑ 1269.image012.png 1270.image002.png 1272.image001.png 1273.image003.png

.image005.png 1275.image002.png 1276.image004.png 1277.image005.png 1278.image007.png 1279.image004.png

.image006.png 1281.image010.png 1282.image007.png 1283.image001.png 1284.image003.png 1285.image002.png

.image004.png 1287.image006.png 1288.image008.png 1289.image001.png 1290.image003.png 1291.image005.png

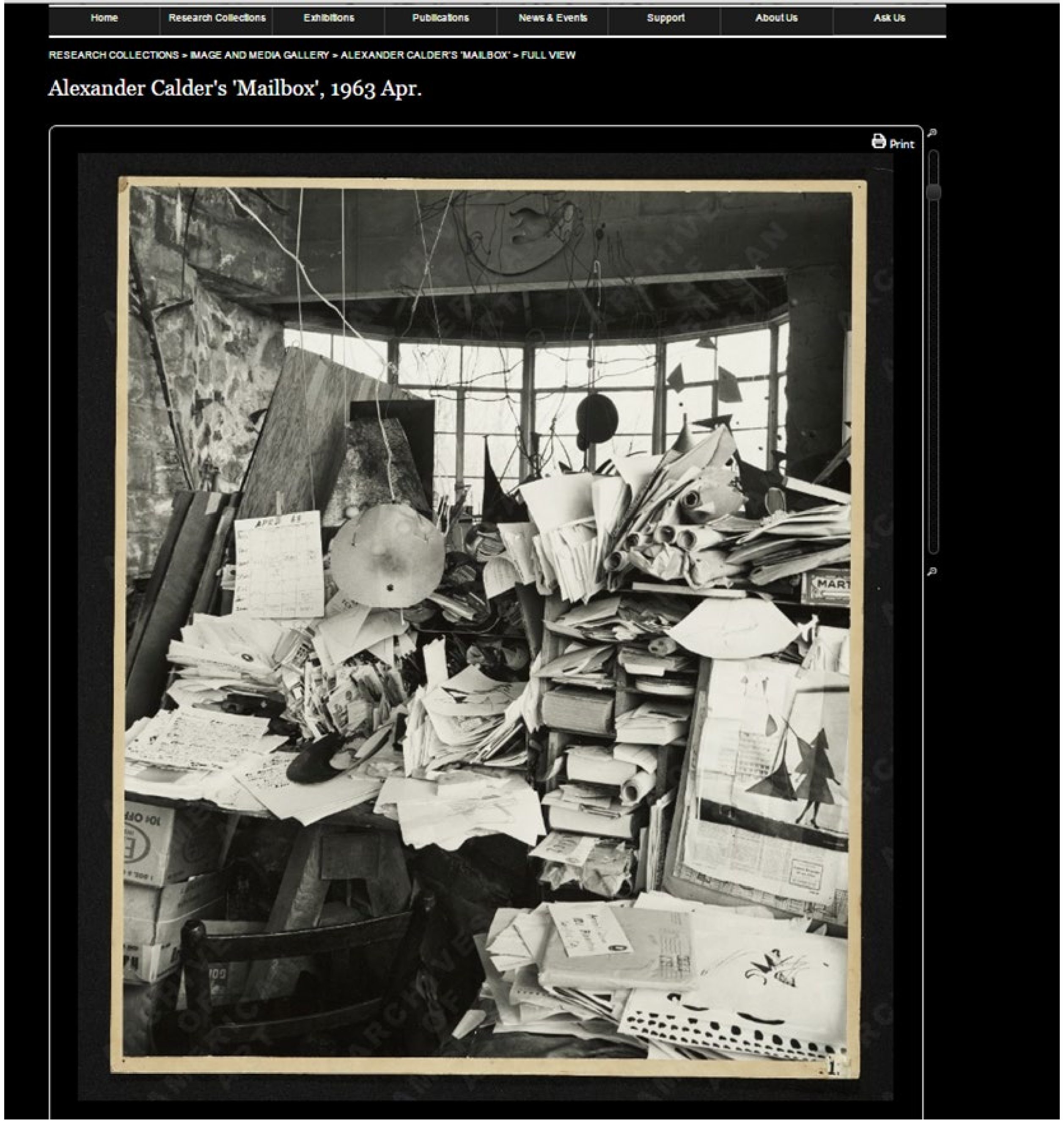
.image004.png 1293.image001.png 1294.image004.png 1295.image006.png 1296.image008.png 1297.image007.png

Smithsonian
Libraries and Archives

Thank you!

Questions?
schmitzfuhrigl@si.edu
or @LyndaLSF

Smithsonian
Libraries and Archives


Home
Research Collections
Exhibitions
Publications
News & Events
Support
About Us
Ask Us
RESEARCH COLLECTIONS > IMAGE AND MEDIA GALLERY > ALEXANDER CALDER'S 'MAILBOX' > FULL VIEW
Print

Alexander Calder's 'Mailbox', 1963 Apr.

Archives of American Art,
(DSI-AAA)6591